The remote sensing of Papyrus vegetation (*Cyperus papyrus L.*) in swamp wetlands of South Africa

Elhadi Mohammed I. Adam

A thesis submitted to the Faculty of Science and Agriculture, at the University of KwaZulu-Natal, in fulfillment of the academic requirements for the degree of Doctor of Philosophy in Environmental Sciences

August 2010

Pietermaritzburg

South Africa

Abstract

Papyrus (*Cyperus papyrus .L*) swamp is the most species rich habitat that play vital hydrological, ecological, and economic roles in central tropical and western African wetlands. However, the existence of papyrus vegetation is endangered due to intensification of agricultural use and human encroachment. Techniques for modelling the distribution of papyrus swamps, quantity and quality are therefore critical for the rapid assessment and proactive management of papyrus vegetation. In this regard, remote sensing techniques provide rapid, potentially cheap, and relatively accurate strategies to accomplish this task.

This study advocates the development of techniques based on hyperspectral remote sensing technology to accurately map and predict biomass of papyrus vegetation in a high mixed species environment of St Lucia- South Africa which has been overlooked in scientific research. Our approach was to investigate the potential of hyperspectral remote sensing at two levels of investigation: field level and airborne platform level.

First, the study provides an overview of the current use of both multispectral and hyperspectral remote sensing techniques in mapping the quantity and the quality of wetland vegetation as well as the challenges and the need for further research.

Second, the study explores whether papyrus can be discriminated from each one of its co-existence species (binary class). Our results showed that, at full canopy cover, papyrus vegetation can be accurately discriminated from its entire co-existing species using a new hierarchical method based on three integrated analysis levels and field spectrometry under natural field conditions. These positive results prompted the need to test the use of canopy hyperspectral data resampled to HYMAP resolution and two machine learning algorithms in identifying key spectral bands that allowed for better discrimination among papyrus and other co-existing species (n = 3) (multi-class classification). Results showed that the random forest algorithm (RF) simplified the process by identifying the minimum number of spectral bands that provided the best overall accuracies. Narrow band NDVI and SR-based vegetation indices calculated from hyperspectral data as well as some vegetation indices published in literature were investigated to test their potential in improving the classification accuracy of wetland plant species. The study also evaluated the robustness and reliability of RF as a variables selection method and as a classification algorithm in identifying key spectral bands that allowed for the successful classification of wetland species.

Third, the focus was to upscale the results of field spectroscopy analysis to airborne hyperspectral sensor (AISA eagle) to discriminate papyrus and it co-existing species. The results indicated that specific wavelengths located in the visible, red-edge, and near-infrared region of the electromagnetic spectrum have the highest potential of discriminating papyrus from the other species.

Finally, the study explored the ability of narrow NDVI-based vegetation indices calculated from hyperspectral data in predicting the green above ground biomass of papyrus. The results demonstrated that papyrus biomass can be modelled with relatively low error of estimates using a non-linear RF regression algorithm. This provided a basis for the algorithm to be used in mapping wetland biomass in highly complex environments.

Overall, the study has demonstrated the potential of remote sensing techniques in discriminating papyrus swamps and its co-existing species as well as in predicting biomass. Compared to previous studies, the RF model applied in this study has proved to be a robust, accurate, and simple new method for variables selection, classification, and modelling of hyperspectral data. The results are important for establishing a baseline of the species distributions in South African swamp wetlands for future monitoring and control efforts.

Preface

The research work described in this thesis was carried out in the School of Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, from August 2006 to June 2010, under the supervision of Professor Onisimo Mutanga (School of Environmental Sciences, University of KwaZulu-Natal; UKZN, South Africa)

I would like to declare that the research work reported in this thesis has never been submitted in any form for any degree or diploma to any tertiary institution. It, therefore, represents my original work .Where use has been made of the work of other authors or organizations it is duly acknowledged within the text or references chapter.

Elhadi Mohammed I. Adam ________________________________Date:________________

As the candidate's supervisor, I certify the above statement and have approved this thesis for submission.

1. Prof. Onisimo Mutanga Signed: __________________________ Date: ________________

Declaration 1-Plagiarism

I, Elhadi Mohammed I. Adam, declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs, or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

   a. Their words have been re-written, but the general information attributed to them has been referenced.

   b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain text, graphics, or tables copied and pasted from the Internet, unless specifically acknowledged and the source being detailed in the thesis and in the References section.

Signed________________________________

Declaration 2- Publication and manuscripts

1. **Adam,** E., Mutanga, O., and Rugege, D. (2010). Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation. *Wetland Ecology and Management*, 18, 281-296.
2. **Adam,** E. and Mutanga, O. (2009). Spectral discrimination of papyrus vegetation (Cyperus papyrus L.) in swamp wetland using field spectrometry. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64, 612-620.
3. **Adam,** E., Mutanga, O., Rugege, D., and Ismail, R. (2009). Field spectrometry of papyrus vegetation (Cyperus papyrus L.) in swamp wetlands of St Lucia, South Africa. *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, IV-260 – IV-263.
4. **Adam,** E., Mutanga, O., Rugege, D., and Ismail, R. (In press). Discriminating the papyrus vegetation (*Cyperus papyrus L.*) and its co-existent species using RF and hyperspectral data resampled to HYMAP. *International Journal of Remote Sensing*.
5. **Adam,** E. and Mutanga, O. (In review). Improving the spectral discriminating of papyrus (*Cyperus papyrus L.*) and its co-existent species at canopy level with hyperspectral indices and RF algorithm. *International Journal of Applied Earth Observation and Geoinformation.*
6. **Adam,** E. and Mutanga, O. (In preparation). Mapping papyrus vegetation (*Cyperus papyrus L.*) and its co-existent species using airborne hyperspectral imagery and RF algorithm.
7. **Adam,** E. and Mutanga, O. (In review). Estimating papyrus (*Cyperus papyrus*) biomass using narrow band vegetation indices and the random forest RF regression algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing.*

Signed________________________________

Dedication

To my dearly loved parents Mohammed Ibrahim and Fatima Abdallah, my much-loved wife, Waheeba Madani, and precious sons, Saeb and Dhafir, for their constant support and prayers for my success

To the departed soul of my brother, Sharief, may Allah forgive him and grant him Paradise.

Acknowledgements

First and foremost, I thank Allah (SWT) for granting me power, support and much more to make it possible to complete my PhD thesis.

I thank Elfashir University for giving me a scholarship to read for a PhD. Special Thank goes to Dr. Mahadi Abdullah Hassan (University of Juba) for his continued encouragement and facilitating the administrative procedures of my scholarship. My appreciation extends to the commitment of the Department of Science and technology (DST)-South Africa in offering me hyperspectral image acquisition in the Dukuduku forest. Without this image, my study could not have been accomplished.

I am all full of admiration and gratitude to Prof Onisimo Mutanga, my supervisor and promotor, for his confidence, scientific guidance, commitment, critical comments, and moral support which enable me to complete my PhD research. This would not have been possible without his encouragement. He taught me how to be an independent scientist, and I learnt how to critically and scientifically write and review research work that saw our papers being readily accepted by the international journals. I always appreciated his coming over to the study area and sloshing through wetland and getting his feet wet to discuss and to develop ideas for my research.

I thank my other promoter, Dr Denis Rugege, who left the UKZN before I completed my study, for teaching me how to do science. Together with Prof. Mutanga, we accomplished the most difficult part of my research: that was writing the research proposal and performing the first field work. I won’t forget the time that we spent together with Prof. Mutanga in the field in the Greater St Lucia Wetlands Park and you continuously advice us to be careful of the wildlife and crocodiles. From your explanations in the field, I could better understand how to identify the vegetation species and the ecological and hydrological processes taking place in the wetland.

My appreciation extends to Dr Moses Cho (CSIR- Council for Scientific and Industrial Research) for helping in the acquisition and pre-processing of the imagery, and without this hyperspectral image, critical scientific objectives could not have been achieved.

My gratitude goes to R Development Core Team for their very powerful open source packages for statistical analysis. Special thanks go to Dr Riyad Ismail (Sappi forests, South

Africa and UKZN research fellowship) for his steadfast support and critical comments. Riyad, you introduced me into the world of statistical script writing in R. I was always hesitant to give you my papers to review, but your comments helped to polish and sharpen my papers. I won't forget spending days and nights thinking about writing R script for my analysis.

I extend my gratitude to the Scientific Board of the St Lucia Wetlands Park and Dukuduku forest for providing all the necessary support. Riky Taylor is thanked for helping with excellent field logistics. Nerosha Govender is thanked for the hospitality and providing excellent conditions for the fieldwork. Sampling in a wetland environment is a difficult and risky task. I am grateful to Khalid Manssour, Mloutswane, and Abd Allah Ibrahim who helped me achieve my goals in the field.

I also thank the staff of the Department of Geography at the University of KwaZulu-Natal for their support and friendship. My special gratitude goes to Mrs. Shanita Ramroop for arranging safe and comfortable transport and accommodation as well as organizing other logistics related to my study. Mr. Donavan DeVos is thanked for his assistance for being a true friend and neighbour during my study. He was always ready to help me. Donavan, your commitment to helping geography students is remarkable. To the whole cartography unit staff, I extend my thanks for your support.

The Sudanese families and colleagues in KwaZulu-Natal-South Africa are thanked for their support. Sheikh Abd Allah Khair and his family made me and my family feel at home in South Africa. We always liked to drive to Verulam because we were always welcomed with a smiling face and warm hospitality. I thank Dr Elfatih Abd-Rahman for his steadfast support and assistance in answering critical scientific questions. Special thanks go to Dr Hafiz Abd Ahmed Abdelgadir, Dr Ibrahim Aldeen, Zakariyyaa Oumar, Clement Adjorlolo, and Victor Bangamwabo who were supportive and helpful during my PhD study. Special thank goes to Dr Ibrahim Elzain Alamin (Elfashir University) for his steadfast support and assistance in taking care of my parents in Sudan during my absent.

My gratitude goes to my family: my mother and father, brothers, sisters, and all relatives for their love and support. I regret being far from you for more than four years.

Finally, I thank my wife Waheeba, for being devoted wife, a true friend, and a supportive colleague during all the hardships that we faced while she accompanied me in South Africa. Without her steadfast support, my PhD would not have been completed. My achievement is her

reward. My sister-in-law, Zahraa, is greatly appreciated for suspending her undergraduate study to travel to South Africa to take care of young Dhafir and Saeb while we were busy studying. She has proved the meanings of sacrifice and selflessness. My sons, Saeb and Dhafir, it was always painful and difficult for me to leave you at home and go to work during the weekends and after hours while you would have liked me to stay and play with you.

Table of Contents

List of figures

List of tables

CHAPTER ONE

General introduction

1.1 Papyrus swamps (*Cyperus papyrus .L*) in African wetlands

Cyperus papyrus .L, commonly called papyrus, belongs to the family Cyperaceae and is one of the most important wetland species that play vital hydrological, ecological, and economic roles in central tropical and western African wetlands. Specifically, papyrus is confined to a belt across equatorial central Africa within the 17° N and 29°S latitudes (Jones and Muthuri, 1985). In South Africa, papyrus co-occurs with some reeds and sedges (e.g. *Phragmites australis, Echinocloa pyramidalis, P. mauritianus, C. dives,* and *Typha capensis*) in open and regularly flooded areas of the Greater St Lucia Wetland Park, KwaZulu-Natal (Dahlberg, 2005; Adam and Mutanga, 2009).

Papyrus swamps are capable of a high standing biomass, accumulating large quantities of nutrients (Gaudet, 1980; Jones and Muthuri, 1985; Kansiime *et al.*, 2005; Boar, 2006), and they are biologically diverse (Denny, 1997), with important landscape functions (Junk, 2003). Several studies in tropical African wetlands have shown the importance of papyrus in hosting habitats for wildlife and bird species (Harper, 1992; Owino and Ryan, 2007) and offering high nutritive grazing for livestock, especially in the dry season (Muthuri and Kinyamario, 1989). Papyrus also has a high capacity to intercept or transform materials moving from catchments to open waters and therefore improving the water quality and soil stabilization (Denny, 1997; Azza *et al.*, 2000). Despite its relative importance, the existence of papyrus vegetation is endangered due to intensification of agriculture and human encroachment in many parts of Africa (Maclean *et al.*, 2006; Owino and Ryan, 2007). In order to understand the spatial distribution of papyrus swamps and to monitor their functions in the landscape, there is a critical need to develop real-time techniques for modelling the spatial distribution and predicting its biomass for the rapid assessment and proactive management of the papyrus swamps. In this regard, the advent of remote sensing, particularly hyperspectral remote sensing, has offered a unique technique to accomplish this task because of its capability to provide rapid, accurate, relatively inexpensive, and near real-time data over large areas (Ozesmi and Bauer, 2002; Schmidt and Skidmore, 2003; Lu, 2006). Consequently, the challenge would be to assess and monitor both the distribution and quantity (biomass) of papyrus species using remote sensing techniques in order to provide the appropriate level of detail and accuracy for detection and mapping purposes. This facilitates a better understanding of the species-quantity interaction in a spatial context.

1.2 Hyperspectral remote sensing

The prefix, hyper, is derived from Greek, *huper,* meaning above, excessive, or an exaggerated amount. The prefix combined with the word "spectral", whose meaning relates to colours, form the word "hyperspectral" (Borengasser *et al.*, 2007). In remote sensing, the term 'hyperspectral' is synonymous with some other terms such as 'spectrometery', 'spectroscopy', 'spectroradiometry', and 'ultraspectral imaging' (Clark, 1999). Spectrometry or spectroradiometry was originally developed from spectro-photometry; 'spectrometry' is a term used in astronomy and is concerned with the measurement of photons as a function of wavelength (Kumar *et al.*, 2001). Spectroscopy is the branch of physics concerned with the interactions between electromagnetic radiation and matter (Kumar *et al.*, 2001). Spectroscopy is the study of light as a function of wavelength that has been absorbed, reflected, or scattered from the materials. The material properties that specify the response of the material at every wavelength are called spectral properties (Suits, 1983). A spectrometer is an optical instrument used for measuring the spectra emanating from natural surfaces in one or more fixed wavelengths in a laboratory, field, aircraft, or satellite (Kumar *et al.*, 2001). In the 1970s, a group of scientists (Knipling, 1970; Hunt, 1977; Swain and Davis, 1978) were able to develop an understanding of spectral properties of rocks, minerals, and vegetation in terms of the underlying quantum mechanical process in relation to the chemistry of the reflecting object. It was concluded that surface properties can possibly be distinguished by measuring the amount of light that reflects from a surface. When an image is constructed from imaging spectrometer data that measure spectra from contiguous image pixels, the terms used are 'imaging spectroscopy', 'imaging spectrometry', or 'hyperspectral imaging'(Clark, 1999).

Hyperspectral imaging is a new technique that has hundreds of narrow continuous spectral bands between 400 nm and 2500 nm, throughout the visible (0.4 nm to 0.7 nm), near-infrared (0.7 nm to 1 nm), and short wave infrared (1nm to 2.5 nm) portions of the electromagnetic spectrum (Vaiphasa *et al.*, 2005; Govender *et al.*, 2009). These contiguous bands and narrow ranges allow for obtaining a spectrum in each position of the large array of the spatial positions so that each single spectral wavelength can be used to make a recognizable image (Figure 1.1) (Clark, 1999; Mutanga, 2004). This greater spectral dimensionality of hyperspectral remote sensing allows for in-depth examination and discrimination of vegetation types that would be lost using other broad band multispectral scanners (Cochrane, 2000; Mutanga *et al.*, 2003;

Schmidt and Skidmore, 2003; Govender *et al.*, 2009). Therefore, it is hypothesised that hyperspectral sensors could help to overcome limitations of spatial and spectral methods when using the broader bands of multispectral scanner systems, such as the mixed pixel problem in mapping vegetation species and the saturation problem in estimating biomass in more dense and high canopy vegetated areas. In this study, two different hyperspectral sensors were used. Measurements were made at field level using the Analytical Spectral Devices (ASD) FieldSpec®3 spectrometer with 2151 spectral bands from 350 nm to 2500 nm and at the airborne platform level using the AISA Eagle sensor with 231 spectral bands from 393 nm to 900 nm.

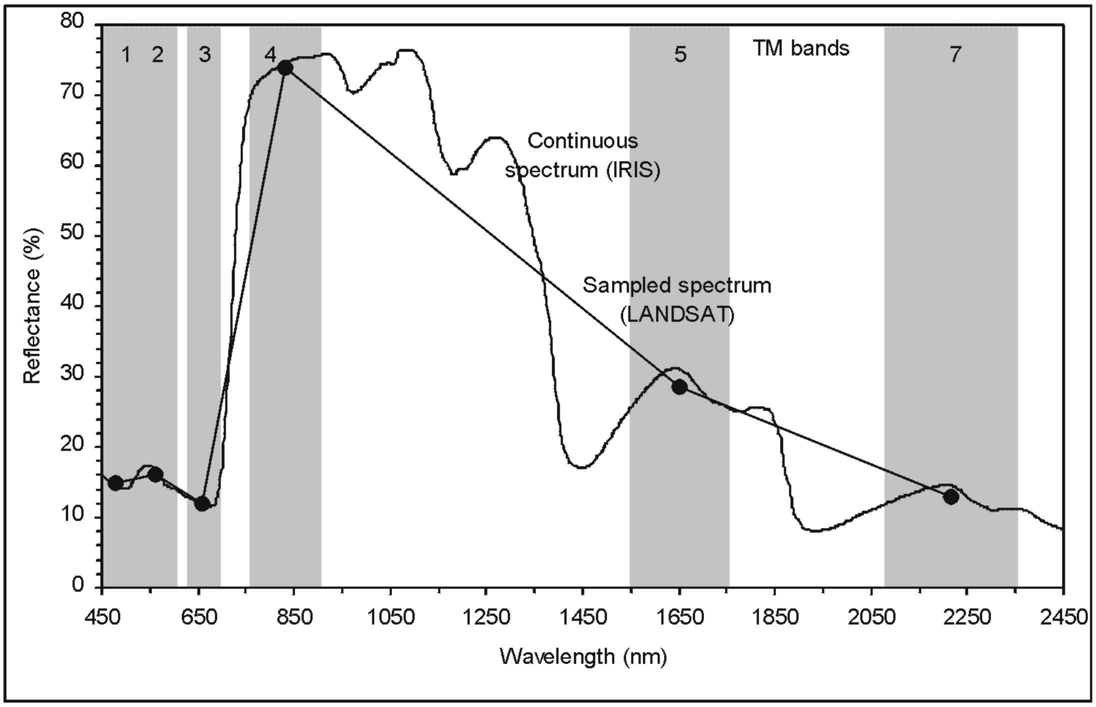

Figure 1.1. A narrow band IRIS (Infrared Intelligent Spectroradiometer) spectrum for fresh green vegetation compared with the discrete wavebands of multispectral LANDSAT TM (Kumar *et al.*, 2001).

1.3 Challenges and opportunities: remote sensing of papyrus vegetation

1.3.1 Discriminating papyrus vegetation using hyperspectral data

Papyrus swamps have increasingly been recognized as being the most habitat rich areas that play ecological, hydrological, and economic roles in tropical wetlands in Africa. To sustain these vital functions of papyrus swamps, a comprehensive understanding of species composition and distribution is therefore critical for their rapid assessment and proactive management (Nagendra, 2001; Schmidt and Skidmore, 2003). Traditionally, species discrimination for floristic mapping requires intensive fieldwork, including taxonomical information, collateral and ancillary data analysis, and the visual estimation of percentage cover for each species. This method is labour-intensive, time-consuming, expensive, and sometimes inapplicable due to the poor accessibility of papyrus swamps and is thus, practical only in relatively small areas. In this context, remote sensing techniques provide rapid, potentially cheap, and relatively accurate strategies for monitoring species composition and distribution.

However, wetland plant species, such as papyrus, are not as easily detectable as terrestrial plant species. This is for two reasons. First, herbaceous wetland vegetation generally exhibits high spectral and spatial variability because of the steep environmental gradients that produce short ecotones and sharp demarcations between the vegetation units (Schmidt and Skidmore, 2003; Adam and Mutanga, 2009; Zomer *et al.*, 2009). Hence, it is often difficult to identify the boundaries between vegetation community types. Second, the reflectance spectra of wetland vegetation canopies are often very similar and are combined with the reflectance spectra of the underlying soil, hydrologic regime, and atmospheric vapour (Guyot, 1990; Malthus and George, 1997; Yuan and Zhang, 2006). This combination usually complicates optical classification and results in a decrease in the spectral reflectance, especially in the near-to mid-infrared regions where water absorption is relatively stronger (Fyfe, 2003; Silva *et al.*, 2008).Therefore, the broad band satellites such as Landsat TM and SPOT, with respect to the sharp ecological gradient with narrow vegetation units in wetland ecosystems, have proven insufficient for discriminating vegetation species in detailed wetland environments (May *et al.*, 1997; Harvey and Hill, 2001; McCarthy *et al.*, 2005).

A significant step forward for remote sensing was made with the development of imaging spectrometry and/or hyperspectral sensors. This development in imaging spectrometry allowed for significant improvement in the accurate detection of small wetland vegetation unit at species level (Daughtry and Walthall, 1998; Schmidt and Skidmore, 2003; Vaiphasa *et al.*, 2005). However, even with the spectral and spatial capabilities of hyperspectral imaging to discriminate between species, studies have shown that the reflectances of vegetation species are highly correlated because of their similar biochemical and biophysical properties (Portigal *et al.*, 1997).

Furthermore, these properties are directly influenced by environmental factors and, therefore, the possibility of a unique spectral signature of a plant species has become questionable (Price, 1994). In addition, spectral variations can also occur within a species because of age differences, micro-climate, soil and water background, precipitation, topography, and stresses (Carter, 1994; Portigal *et al.*, 1997; Garcia and Ustin, 2001; Smith *et al.*, 2004).

On the other hand, the high spectral resolution of hyperspectral data comes with the complexity of the high data dimensionality (Bajwa *et al.*, 2004).This redundant data might be problematic in terms of image processing algorithms, an excessive demand for sufficient field samples, high cost, and overfitting when using multivariate statistical techniques (Bajcsy and Groves, 2004; Borges *et al.*, 2007; Mutanga and Kumar, 2007; Vaiphasa *et al.*, 2007). Therefore, it is imperative to identify the optimal bands required for discriminating and mapping wetland species without losing any important information (Bajcsy and Groves, 2004; Vaiphasa *et al.*, 2007). Various univariate and multivariate band reduction techniques have been developed, such as RF, partial least square regressions, classification trees, discriminant analysis, principal component analysis, and artificial neural network. It is, therefore, important to understand the advantages and disadvantages of band reduction techniques and select accordingly. In this context, the challenge would be to explore and test robust methods and techniques for the effective processing and classification of hyperspectral data for better and more accurate detecting and mapping of papyrus swamps. Furthermore, these methods and techniques need to be automated to some degree with limited human intervention to allow for critical evaluation (Soh and Tsatsoulis, 1999).

1.3.2 Assessment of papyrus quantity using hyperspectral data

Papyrus vegetation is increasingly being recognized for its accumulated large quantities of nutrients (Gaudet, 1980) and high standing biomass productivity (Muthuri and Kinyamario, 1989; Jones and Muthuri, 1997; Kansiime *et al.*, 2005). The value of papyrus swamp in tropical wetlands often depends on the status of its productivity. Efficient techniques that can spatially and temporally monitor the stability of papyrus productivity and whether significant changes are taking place in papyrus swamp are, therefore, required. Measuring the biophysical parameters of papyrus vegetation, such as biomass, is important for quantifying the primary production or carbon cycle of the swamp ecosystem (Jones and Muthuri, 1997; Kansiime *et al.*, 2005). Direct field methods for estimating biomass require frequent destructive harvesting (Lu, 2006). Such traditional methods are expensive, time-consuming, labour-intensive, and difficult to implement, especially in such large and inaccessible areas (Lu, 2006). Remote sensing, particularly spectroscopy, offers advanced and effective techniques that can provide the needed protocols for monitoring papyrus biomass.

Based on broad band satellite images, vegetation indices such as Normalized Difference Vegetation Index (NDVI) and Simple Ratio (RS) have been the most successful in quantifying and monitoring wetland productivity over large areas at open canopy scale (Moreau *et al.*, 2003; Rendonga and Jiyuanb, 2004; Proisy *et al.*, 2007).

In spite of these successes, vegetation indices calculated from broad band sensors can be unstable, owing to the underlying soil colour, canopy and leaf properties, and atmospheric conditions (Huete and Jackson, 1988; Todd *et al.*, 1998). Furthermore, NDVI derived from broad band satellite images such as NOAA or Landsat TM asymptotically saturate after a certain biomass density, and measurement accuracy drops considerably (Tucker, 1977; Gao *et al.*, 2000; Thenkabail *et al.*, 2000). Figure 1.2 shows a hypothetical illustration of this biomass-NDVI relationship.

More recently, the appearance of hyperspectral sensors has opened new perspectives for developing vegetation indices (VIs) using the provided additional narrow bands within the visible, NIR, and short wave infrared (SWIR) with less than 10 nm bandwidths from visible to SWIR (350 nm – 2500 nm) rather than focusing on the red and NIR broad band (Hansen and Schjoerring, 2003; Mutanga and Skidmore, 2004a; Cho *et al.*, 2007; Fava *et al.*, 2009). The use of NDVI calculated from narrow bands has been found to be one possibility for overcoming or reducing the data saturation problem (Mutanga and Skidmore, 2004a). This capability of VIs calculated from narrow bands needs to be tested or improved for better estimation of papyrus biomass in more densely vegetated and wetland areas. In this thesis, it is hypothesised that hyperspectral remote sensing, with its capability to resolve detailed spectral features, can estimate papyrus biomass accurately.

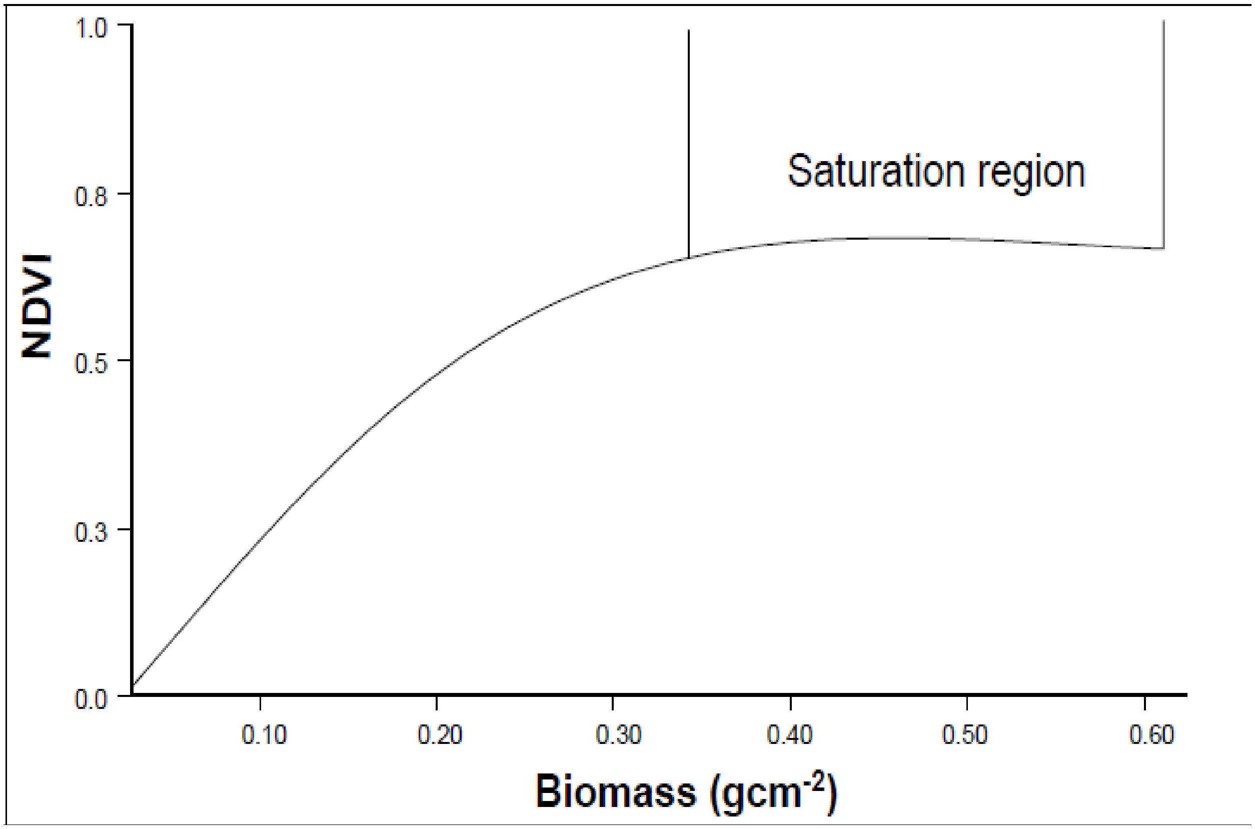

Figure 1.2. Relationship between NDVI and biomass. The saturation level is usually reached at about 0.3 g cm $^{-2}$ (Mutanga, 2004).

1.4 Research objectives

The main aim of this study was to investigate the potential of hyperspectral remote sensing to discriminate papyrus vegetation from its co-existing species and to estimate biomass of papyrus at high canopy density or full canopy level in the Greater St Lucia Wetland Park, South Africa. The specific objectives in this study are as follows:

1. To explore the usefulness of *in situ* spectroscopic data in discriminating papyrus vegetation from its co-existing species (binary class techniques),
2. To investigate the usefulness of *in situ* spectroscopic data in discriminating among papyrus vegetation and its co-existing species (multi-class techniques),
3. To determine if machine learning algorithms (RF) can accurately discriminate among papyrus and other co-existing species using resampled HYMAP data,
4. To examine whether vegetation indices derived from spectroscopy data can be used to enhance the separability and classification accuracy between vegetation species,

5. To test the reliability and robustness of the internal accuracy assessment of the RF as a variable selection and classification algorithm in discriminating between the species,
6. To investigate the potential of imaging spectroscopy in discriminating among papyrus and its co-existing species using airborne hyperspectral data (AISA eagle), and
7. To explore the potential of hyperspectral data in estimating biomass of papyrus at high canopy density or full canopy levels.

1.5 Scope of the study

In this study, the potential use of hyperspectral remote sensing techniques to discriminate and estimate biomass of papyrus swamps in the Greater St Lucia Wetland, South Africa was investigated. Two classification methods were investigated to discriminate papyrus from its co-existing species: binary class which focused on discriminating papyrus from each of its co-existing species and multi-class for discriminating among papyrus and its co-existing species. The use of hyperspectral remote sensing techniques in estimating papyrus biomass was subsequently evaluated.

Two hyperspectral levels were investigated: at field level using a hand-held spectrometer data and at airborne platform level using AISA eagle data. In a follow-up study, the usefulness of hyperspectral data was also evaluated for estimating papyrus biomass at full canopy level. In this context, relatively more emphasis was placed on the prediction of papyrus biomass because it is considered as the most limiting factor for the ecological, hydrological, and economic roles of papyrus in a wetland ecosystem (Muthuri and Kinyamario, 1989; Jones and Muthuri, 1997; Kansiime *et al.*, 2005). The Greater St Lucia Wetland Park (GLWP) in South Africa was used as a test site both for field and airborne spectrometry.

1.6 The study area

The Greater St Lucia Wetland Park is a protected area located on the eastern coast of KwaZulu-Natal Province, about 245 kilometres north of Durban, South Africa. The park stretches from the southern Mozambiqucan coastal plain to KwaZulu-Natal, covering about 328 000 hectares between longitudes $32^{o}21^{'}$ E and $32^{o}34^{'}$ E and latitudes $27^{o}34^{'}$ S and 28^{o} $24^{'}$ S (Figure 1.3). Therefore, the GSWP is considered to be the largest estuarine system in Africa (Taylor, 1995).

The climate is sub-tropical with the mean annual rainfall varying from 1500 mm on the eastern shore to 700 mm on the western shore of the lake (Taylor, 1995). The GSWP is characterized by a high diversity of ecosystems including marine, inland lake, estuarine, forested dune, mangrove, and coastal and swamp forest. The area is permanently either wet or flooded with freshwater throughout the year and is recognized as a UNESCO World Heritage Site and a Ramsar wetland of global significance. The park supports extraordinary ecological and biological diversity due to its location that is between tropical and sub-tropical biota.

Different wetland vegetation species cover the park including those in salt marshes (e.g. *Juncus krausii, Salicornia spp., and Ruppia maritima*); Saline reed swamps (*Phragmites mauritianus*); Sedge Swamp (*Eleocharis limosa*), and Echinochloa floodplain grassland (*Echinochloa pyramidalis, Eriochloa spp., and Cyperus spp*.), but the most dominant species are found in the freshwater swamps and are reed and papyrus (*Phragmites australis* and *Cyperus papyrus*) as well as *Echinochloa pyramidalis* and *Thelypteris interrupta*. In this study, four study sites were focused on including Futululu forest, the Dukuduku Indigenous Forest, Mfabeni swamps and Mkuzi swamps (Figure 1.3). At these sites, papyrus (*Cyperus papyrus)* occurs in large areas between forested dunes and plantation forest on organic and alluvial soil with mainly three other species including *Phragmites australis*, *Echinochloa pyramidalis*, and *Thelypteris interrupta* (Adam and Mutanga, 2009).

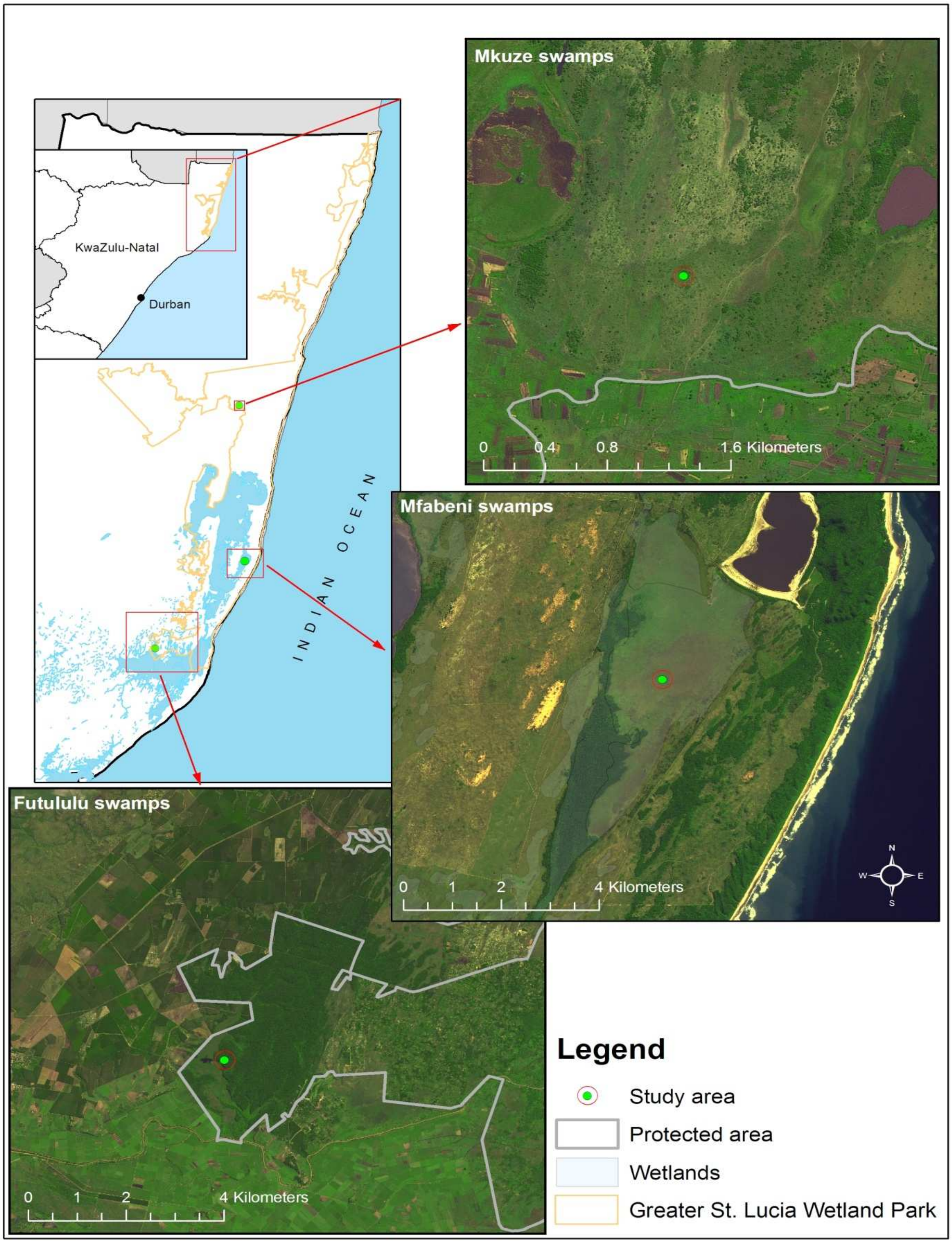

Figure 1.3. Location of the study area in KwaZulu-Natal Province of South Africa.

1.7 Thesis outline

To achieve the main objectives of this study, the thesis is organized as a collection of 6 research papers that have been submitted to peer reviewed international journals. Of these 6 papers, 3 papers have already been published and 2 papers are still in review and the remaining in preparation. Each paper has been written as a stand-alone article that can be read separately from the rest of the thesis but that draws separate conclusions that link to the overall research objectives and questions. As a result, a number of overlaps and replications occur in the sections "Introduction" and "Method" in the different chapters. This problem is deemed to be of little significance when one considers the critical peer review process and the fact that the different chapters are papers that can be read separately without losing the overall context. The thesis consists of 8 chapters:

Chapter 2 contains a detailed literature review of the relevant application of multispectral and hyperspectral remote sensing in discriminating and estimating some of the biophysical and biochemical parameters of wetland vegetation. Specific relevance to the objectives of this study is highlighted in Section 2.6 (spectral discrimination of wetland species using hyperspectral data) and Section 2.7 (estimating biophysical and biochemical parameters of wetland species). The research gaps and challenges in the application of hyperspectral remote sensing in wetland species are introduced.

Chapter 3 contains an investigation into the ability of hyperspectral data to discriminate between papyrus vegetation and its co-existing species. The study determines if there is a significant difference in the mean of reflectance between the pairs of papyrus and each one of the co-existing species (binary class) at each measured wavelength from 350 nm to 2500nm. For the wavelengths that are significantly different (p < 0.001), it was tested whether some wavelengths have more discriminating power than others and which band combinations can yield the lowest misclassification rate.

Chapter 4 contains the findings of an investigation into the potential use of machine learning algorithms (RF) and resampled HYMAP data to accurately discriminate between papyrus and its co-existing species at canopy level. In this chapter, the work presented in Chapter 3 is extended from binary class classification to multi-class classification to assess the use of spectroscopic data in discriminating between papyrus and its co-existing species at canopy level under natural field conditions using RF algorithms and variables selection methods.

Chapter 5 investigates the potential of several vegetation indices derived from hyperspectral data to better improve the discriminating accuracy between papyrus and other species using RF ensembles. Specifically, the study examined the ability of widely used indices (NDVI and SR) calculated from hyperspectral bands to identify the most important portions of the electromagnetic spectrum that could yield high accuracy in discriminating between papyrus and its co-existing species at canopy level. Some vegetation indices published in the literature were also investigated and new indices were proposed.

Chapter 6 is based on the observations and conclusions drawn from Chapter 3 to Chapter 5 to develop the best approach for discriminating between papyrus and its co-existing species using airborne hyperspectral imagery (AISA eagle).

Chapter 7 evaluates the utility of the widely used indices (NDVI and SR) derived from hyperspectral bands to identify the most sensitive regions of the electromagnetic spectrum that could be used to estimate papyrus biomass at high canopy density. The RF regression algorithm was implemented to test whether narrow band vegetation indices could predict papyrus biomass under field conditions.

Finally, a synthesis of the study is provided in Chapter 8. The findings are summarized and conclusions are derived from the preceding chapters. Some relevant recommendations for future research on the applications of remote sensing in wetland vegetation mapping are outlined. A special focus is directed towards the operational use of remote sensing techniques in mapping and monitoring of papyrus swamps.

A single reference list is provided at the end of the thesis.

CHAPTER TWO

Literature review

This chapter is based on:

Adam, E., Mutanga, O. and Rugege, D. (2009). Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation. *Wetland Ecology and Management*, 18,281-296.

Abstract

Wetland vegetation plays a key role in the ecological functions of wetland environments. Remote sensing techniques offer timely, up-to-date, and relatively accurate information for sustainable and effective management of wetland vegetation. This article provides an overview on the status of remote sensing applications in discriminating and mapping wetland vegetation, and estimating some of the biochemical and biophysical parameters of wetland vegetation. Research needs for successful applications of remote sensing in wetland vegetation mapping and the major challenges are also discussed. The review focuses on providing fundamental information relating to the spectral characteristics of wetland vegetation, discriminating wetland vegetation using broad and narrow bands, as well as estimating water content, biomass, and leaf area index. It can be concluded that the remote sensing of wetland vegetation has some particular challenges that require careful consideration in order to obtain successful results. These include an in-depth understanding of the factors affecting the interaction between electromagnetic radiation and wetland vegetation in a particular environment, selecting appropriate spatial and spectral resolution as well as suitable processing techniques for extracting spectral information of wetland vegetation

Keywords: Biomass. Leaf area index**.** Mapping. Remote sensing. Water content. Wetland vegetation**.**

2.1 Introduction

Wetland vegetation is an important component of wetland ecosystems that plays a vital role in environmental function (Kokaly and Clark, 1999a; Yuan and Zhang, 2006). It is also an excellent indicator for early signs of any physical or chemical degradation in wetland environments (Dennison et al., 1993).

Mapping and monitoring vegetation species distribution, quality, and quantity are important technical tasks in sustainable management of wetlands. This task involves a wide range of functions including natural resource inventory and assessment, fire control, wildlife feeding, habitat characterization, and water quality monitoring at a given time or over a continuous period (Carpenter *et al.*, 1999). Moreover, it is essential to have up-to-date spatial information about the magnitude and the quality of vegetation cover in order to initiate vegetation protection and restoration programmes (He *et al.*, 2005).

Traditionally, species discrimination for floristic mapping requires intensive fieldwork, including taxonomical information, collateral and ancillary data analysis, and the visual estimation of percentage cover for each species; this is labour-intensive, costly, and time-consuming and sometimes inapplicable due to the poor accessibility, and is thus, only practical on relatively small areas (Hardisky *et al.*, 1986; Lee and Lunetta, 1995; Klemas, 2001). Remote sensing, on the other hand, offers a practical and economical means to discriminate and estimate the biochemical and biophysical parameters of the wetland species, and it can make field sampling more focused and efficient. Its repeat coverage offers archive data for detection of change over time, and its digital data can be easily integrated into Geographic Information System (GIS) for more analysis (Shaikh *et al.*, 2001; Ozesmi and Bauer, 2002). For this advantage, many researchers have used both multispectral data such as Landsat TM and SPOT imagery to identify general vegetation classes or to attempt to discriminate broad vegetation communities (May *et al.*, 1997; Harvey and Hill, 2001; Li *et al.*, 2005) as well as classify and map wetland vegetation at the species level using hyperspectral data (Schmidt and Skidmore, 2003; Rosso *et al.*, 2005; Vaiphasa *et al.*, 2005; Belluco *et al.*, 2006; Pengra *et al.*, 2007). Moreover, the use of remote sensing techniques has been extended into measuring the biophysical and biochemical properties such as leaf area index (LAI), biomass, and water content of wetland vegetation (Penuelas *et al.*, 1993a; Rendonga and Jiyuanb, 2004; Proisy *et al.*, 2007).

The rapid growth in the number of studies that have investigated the use of remote sensing in studying wetland species makes it necessary to provide an overview of the techniques that have been used and to identify those aspects that still need further investigation. This would be useful practically in wetland management and scientifically through highlighting the priorities and challenges for further research.

Previous reviews on remote sensing of wetlands included those by Silva *et al.* (2008) who discussed the theoretical background and applications of remote sensing techniques in aquatic plants in wetland and coastal ecosystems. Ozesmi and Bauer (2002) reviewed the classification techniques used to map and delineate different wetland types using different remotely sensed data. Lee and Lunetta (1995) reviewed the use and the cost of airborne and satellite sensors in the inventory of and change detection in wetlands. The review by Klemas (2001) addressed the current use of remote sensing and its opportunities pertinent in monitoring the environmental indicators in coastal ecosystems. Hardisky *et al.* (1986) reviewed different remotely sensed data for coastal wetlands and estimating biomass.

The limitation of the above-mentioned reviews is that no specific aspect of the application of remote sensing has been addressed individually and most of the reviews have been focused on the use of remote sensing in mapping and identification of wetland types at a broad level. There has been no specific review on the use of both hyperspectral and multispectral remote sensing in discriminating wetland vegetation as well as estimating its biophysical and biochemical properties which is essential in wetland management. Hence, this review focuses specifically on the application of remote sensing in discriminating and estimating the biophysical and biochemical properties of wetland vegetation.

The specific objectives of this study were to review the status of application of both multispectral and hyperspectral remotely sensed data in wetland vegetation with special focus on: 1. discriminating and mapping wetland vegetation, 2. estimating some of the biophysical and biochemical properties of wetland vegetation, and 3. highlighting the major challenges and further research needed for a successful application of remote sensing in wetland vegetation.

2.2 Challenges in mapping wetland vegetation

Wetland plants and their properties are not as easily detectable as terrestrial plants, which occur in large stratification. This is for two reasons. First, herbaceous wetland vegetation exhibits high spectral and spatial variability because of the steep environmental gradients which produce short ecotones and sharp demarcation between the vegetation units (Schmidt and Skidmore, 2003; Adam and Mutanga, 2009; Zomer *et al.*, 2009). Hence, is often difficult to identify the boundaries between vegetation community types. Second, the reflectance spectra of wetland vegetation canopies are often very similar and are combined with reflectance spectra of the underlying soil , hydrologic regime, and atmospheric vapour (Guyot, 1990; Malthus and George, 1997; Yuan and Zhang, 2006). This combination usually complicates the optical classification and results in a decrease in the spectral reflectance, especially in the near-to mid-infrared regions where water absorption is stronger (Figure 2.1) (Fyfe, 2003; Silva *et al.*, 2008).Therefore, the current efforts which have been successful at mapping terrestrial vegetation using optical remote sensing, may not be able, either spatially or spectrally, to effectively distinguish the flooded wetland vegetation because the performance of near- to mid-infrared bands are attenuated by the occurrences of underlying water and wet soil (Schmidt and Skidmore, 2003; Hestir *et al.*, 2008). However, hyperspectral narrow spectral channels offer the potential to detect and map the spatial heterogeneity of wetland vegetation (Schmidt and Skidmore, 2003; Vaiphasa *et al.*, 2007; Hestir *et al.*, 2008).

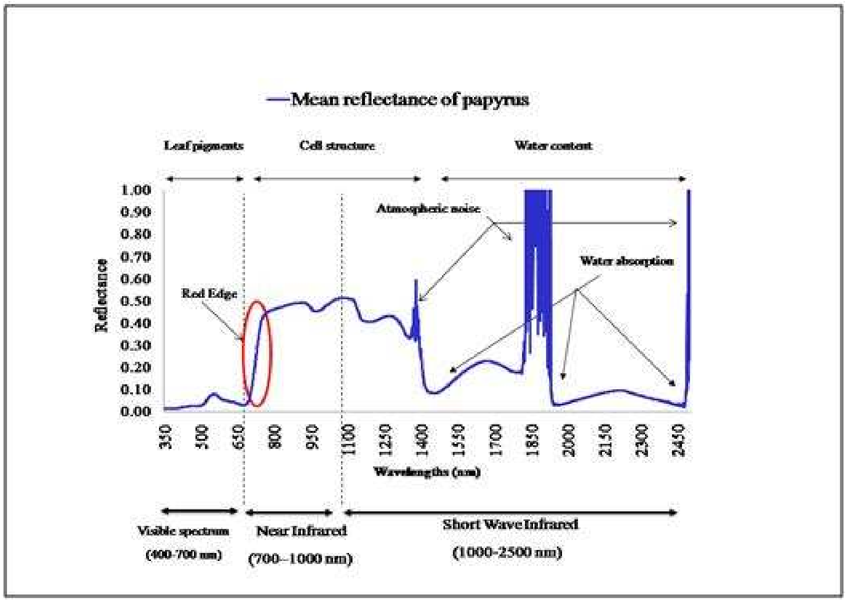

Figure 2.1. Mean canopy reflectance spectra of *Cyperus papyrus L.* in swamp wetland with the dominating factor influencing each interval of the curve. Most of the short wave infrared wavelengths (water content wavelength) are affected by atmospheric noise.

2.3 Factors affecting spectral characteristics of wetland vegetation

When solar radiation interacts with leaves, it may be reflected, absorbed, and/or transmitted. All vegetation species contain the same basic components that contribute to its spectral reflectance, including chlorophyll and other light-absorbing pigments, water, proteins, starches, waxes, and structural biochemical molecules, such as lignin and cellulose (Price, 1992; Kokaly and Clark, 1999b). Hence, the spectral separability of vegetation species is challenging due to those limiting factors affecting the spectral response of vegetation species (Price, 1992; Rosso *et al.*, 2005). In general, the spectral differences among vegetation species are normally derived from leaf optical properties related to the biochemical and biophysical status of the plants. Leaf optical properties depend on leaf surface and internal structure, leaf thickness, water content, biochemical composition, and pigment concentration (Kumar *et al.*, 2001; Rosso *et al.*, 2005). The spectral reflectance of wetland vegetation is normally subdivided into four domains. While vegetation types generally have a high reflectance and transmittance in the near-infrared region and strong water absorption in the mid-infrared region (Figure 2.1), the spectral reflectance of wetland vegetation is normally divided into four domains as shown in Table 2.1.

Table 2.1: The spectral reflectance of green vegetation on the four regions of electromagnetic spectrum defined by Kumar *et al*. (2001)

Wavelengths region (nm)	description	Spectral reflectance of vegetation	References
400-700	Visible	Low reflectance and transmittance due to chlorophyll and carotene absorption	(Kumar *et al.*, 2001; Rosso *et al.*, 2005)
680-750	Red-edge	The reflectance is strongly correlated with plant biochemical and biophysical parameters.	(Clevers, 1999; Mutanga and Skidmore, 2007)
700-1300	Near-infrared	High reflectance and transmittance, very low absorption. The physical control is internal leaf structures.	(Kumar *et al.*, 2001; Rosso *et al.*, 2005)
1300-2500	Mid-infrared	Lower reflectance than other spectrum regions due to strong water absorption and minor absorption of biochemical content.	(Kumar *et al.*, 2001)

The most important factors affecting the spectral reflectance among wetland vegetation are the biochemical and biophysical parameters of the plants' leaves and canopy such as chlorophyll a and b, carotene, and xanthophylls (Guyot, 1990; Kumar *et al.*, 2001). Wetland species appear to vary greatly in chlorophyll and biomass reflectance as a function of plant species and hydrologic regime (Anderson, 1995). Spectral behaviour of wetland vegetation is also influenced by leaf water content which determines the absorption of the mid-infrared region (Datt, 1999). Red reflectance increases with leaf water stress through an association with a reduction in chlorophyll concentration (Filella and Penuelas, 1994).The relationship between the increase of near-infrared leaf reflectance and decrease of leaf water content has also been reported (Aldakheel and Danson, 1997). For example, Yuan and Zhang (2006) compared the laboratory and field spectral characteristics of the submerged plant (Vallisneria *spiralis*) in the constructed wetland at Shanghai in China. They found that the spectral reflectance measured by the ground-based spectroradiometer sensor was a combination of plant spectra, segmental water, and fundus spectrum.

Leaf area index is also a key variable in the canopy reflectance of the wetland vegetation. The canopies with a high LAI reflect more than the canopies with medium or low LAI. However, higher LAI canopies allow only little light radiation to reach to the mature leaves under vegetation canopies and the soil background (Abdel-Rahman and Ahmed, 2008; Darvishzadeh *et al.*, 2008). Studies show that the spectral signature of tropical wetland canopies is also affected by the different seasons, plant architecture, and illumination angle (Cochrane, 2000; Artigas and Yang, 2005; Darvishzadeh *et al.*, 2008).

2.4 Mapping wetland vegetation using multispectral data

Historically, aerial photography was the first remote sensing method to be employed for mapping wetland vegetation (Seher and Tueller, 1973; Shima *et al.*, 1976; Howland, 1980; Lehmann and Lachavanne, 1997). These studies concluded that aerial photography is most useful for detailed wetland mapping because of its minimum mapping unit (MMU). However, aerial photography is not feasible for mapping and monitoring wetland vegetation on a regional scale or for monitoring that requires continual validation of information because it is costly and time-consuming to process.

Currently, a variety of remotely sensed images are available for mapping wetland vegetation at different levels by a range of airborne and spaceborne sensors from multispectral sensors to hyperspectral sensors which operate within the different optical spectra, with different spatial resolutions ranging from sub-metre to kilometres and with different temporal frequencies ranging from 30 minutes to weeks or months. Among them, aerial photography, Landsat TM, and SPOT images were commonly investigated in mapping vegetation types in wetlands. The common image analysis techniques used in mapping wetland vegetation include digital image classification (i.e. unsupervised and supervised classification) (May *et al.*, 1997; Harvey and Hill, 2001; McCarthy *et al.*, 2005) and vegetation index clustering (Nagler *et al.*, 2001; Yang, 2007). May *et al.* (1997) compared Landsat TM and SPOT multispectral data in mapping shrub and meadow vegetation in northern California. They concluded that Landsat TM data were more effective than SPOT data in separating shrubs from meadows. However, neither Landsat TM nor SPOT data were effective in distinguishing meadow sub-types. McCarthy *et al.* (2005) in Botswana found that the high spatial and temporal variation in vegetation in the Okavango Delta makes ecoregion classification from Landsat TM data unsatisfactory for achieving land cover classification. In Australian wetlands, Landsat TM has proven to be a potential source of defining vegetation density, vigour, and moisture status, but not efficient in defining the species composition (Johnston and Barson, 1993). Harvey and Hill (2001) in the Northern Territory, Australia, compared aerial photographs, SPOT XS, and Landsat TM image data to determine the accuracy and applicability of each data source for the spectral discrimination of vegetation types. Their results demonstrated that aerial photography was clearly superior to SPOT XS and Landsat TM imagery for detailed mapping of vegetation communities in the tropical wetland. They also found that the sensitivity of Landsat band 2 (green), band 3 (red), band 4 (near-infrared, NIR), and band 5 (mid-infrared, MIR) provided a more accurate classification than SPOT. Ringrose *et al*. (2003) used NOAA-AVHRR and SPOT to map the ecological conditions at the Okavango delta in Botswana. They concluded that it was difficult to discriminate grassed floodplain from wooded peripheral drylands. Sawaya *et al.* (2003) at Minnesota in USA were able to map the vegetation groups at a local scale using IKONOS imagery with a high level of classification accuracy (80%).

Imagery from the Landsat TM and SPOT satellite instruments have proven insufficient for discriminating vegetation species in detailed wetland environments (May *et al.*, 1997; Harvey and Hill, 2001; Ringrose *et al.,* 2003; Sawaya *et al.,* 2003; McCarthy *et al.,* 2005).This is due to: 1. the difficulties faced in distinguishing fine, ecological divisions between certain vegetation species, 2. the broad nature of the spectral wavebands with respect to the sharp ecological gradient with narrow vegetation units in wetland ecosystems, and 3. the lack of high spectral and spatial resolution of optical multispectral imagery which restricts the detection and mapping of vegetation types beneath a canopy of vegetation, in densely vegetated wetlands.

Although these studies produced reasonable results on mapping wetland vegetation at a regional scale and vegetation communities, more research is needed to explore the benefits of incorporating bathymetric and other auxiliary data to improve the accuracy of mapping wetland vegetation at the species level.

2.5 Improving the accuracy of wetland vegetation classification

Spectral discrimination between vegetation types in complex environments is a challenging task, because commonly different vegetation types may possess the same spectral signature in remotely sensed images (Domaç and Süzen, 2006; Sha *et al.*, 2008; Xie *et al.*, 2008). Traditional digital imagery from multispectral scanners is subject to limitations of spatial, spectral, and temporal resolution. Moreover, applications of per-pixel classifiers to images dominated by mixed pixels are often incapable of performing satisfactorily and produce inaccurate classification (Zhang and Foody, 1998). Due to the complexities involved, more powerful techniques have been developed to improve the accuracy of discriminating vegetation types in remotely sensed data.

Domaç and Süzen (2006) in the Amanos Mountains region of southern-central Turkey used knowledge-based classifications in which they combined Landsat TM images with environmental variables and forest management maps to produce regional scale vegetation maps. They were able to produce an overall high accuracy when compared with the traditional maximum likelihood classification method. Another example for improving classification accuracy by incorporating vegetation-related environmental variables using GIS with remotely sensed data was the work of Yang (2007) at Hunter Region in Australia. He used digital aerial photographs, SPOT-4, and Landsat-7 ETM+ images for riparian vegetation delineation and mapping. The overall vegetation classification accuracy was 81% for digital aerial photography, 63% for SPOT-4, and 53% for Landsat-7 ETM+. The study revealed that the lack of spectral resolution of aerial photographs and the coarse spatial resolution of the current satellite images is the major limiting factor for their application in wetland vegetation mapping.

Artificial neural network (ANN) and fuzzy logic approaches were also investigated to improve the accuracy of mapping vegetation types in complex environments. ANN proved to be valuable in mapping vegetation types in wetland environments. One disadvantage of ANN, however, is that ANN can be computationally demanding to train the network when large datasets are dealt with (Carpenter *et al.*, 1999; Berberoglu *et al.*, 2000; Filippi and Jensen, 2006; Xie *et al.*, 2008). Berberoglu *et al.* (2000) at the Cukurova Deltas in Turkey combined ANN and texture analysis on a per-field basis to classify land cover from Landsat TM. They were able to increase the accuracy achieved with maximum likelihood classification by 15%. Carpenter *et al.* (1999) compared conventional expert methods and the ARTMAP neural network method in mapping vegetation types at the Sierra National Forest in Northern California using Landsat TM data. Their research illustrated that the accuracy was improved from 78% in conventional expert methods to 83% when the ARTMAP neural network method was used. The ARTMAP neural network method was found to be less time-consuming and its production to be easily updated with any new observation.

A fuzzy classification technique, which is a kind of probability-based classification rather than a crisp classification, is also useful in mixed-class areas and was investigated for solving the problem of mapping complex vegetation. Sha *et al*. (2008) at the Xilinhe River Basin in China employed a hybrid fuzzy classifier (HFC) for mapping vegetation on typical grassland using Landsat ETM+ imagery. It was concluded that HFC was much better than conventional supervised classification (CSC) with an accuracy percentage of 80.2% as compared to 69.0% for the CSC. Promising results have also been achieved in using fuzzy classification for suburban land cover classification from Landsat TM and SPOT HRV data by Zhang and Foody (1998) at Edinburgh in Scotland. They concluded that fuzzy classification not only has advantages over conventional hard methods and partially fuzzy approaches, but also is more feasible in integrating remotely sensed data and ancillary data.

Decision tree (DT) classification has also shown promising results in mapping vegetation in wetlands and complex environments. DT is a simple and flexible non-parametric rule-based classifier and it can handle data that are represented on different measurement scales. This is useful especially when there is a need to integrate the environmental variables (e.g. slope, soil type, and rainfall) in the mapping process (Xu *et al.*, 2005; Xie *et al.*, 2008). Xu *et al.* (2005) at

Syracuse in New York employed a decision tree and regression (DTR) algorithm to determine class proportions within a pixel so as to produce soft land cover classes from Landsat ETM. Their results clearly demonstrate that DTR produces considerably higher soft classification accuracy (74.45%) as compared to the conventional maximum likelihood classifier (MLC) (55.25%) and the fuzzy C-means supervised (FCM) (54.40%).

It has been revealed from the present review that no single classification algorithm can be considered as an optimal methodology for improving vegetation discrimination and mapping. Hence, the use of advanced classifier algorithms must be based on their suitability to achieve certain objectives in specific areas.

2.6 Spectral discrimination of wetland species using hyperspectral data

In remote sensing, the term ‘imaging spectroscopy’ is synonymous with some other terms such as ‘imaging spectrometry’ and ‘hyperspectral’ or ‘ultraspectral imaging’ (Clark, 1999). In general, hyperspectral remote sensing has hundreds of narrow continuous spectral bands between 400 nm and 2500 nm, throughout the visible (0.4 nm to 0.7 nm), near-infrared (0.7 nm to 1 nm), and short wave infrared (1nm to 2.5 nm) portions of the electromagnetic spectrum (Vaiphasa *et al.*, 2005; Govender *et al.*, 2009). This greater spectral dimensionality of hyperspectral remote sensing allows in-depth examination and discrimination of vegetation types which would be lost with other broad band multispectral scanners (Cochrane, 2000; Mutanga *et al.*, 2003; Schmidt and Skidmore, 2003; Govender *et al.*, 2009). Hyperspectral remote sensing data is mostly acquired using a hand-held spectrometer or airborne sensors. A hand-held spectrometer is an optical instrument used for measuring the spectrum emanating from a target in one or more fixed wavelengths in the laboratory and the field (Kumar *et al.*, 2001). The accurate measurements of the spectral reflectance in the field were established in the 1960s as a result of the rapid growth in airborne multispectral scanners (Milton *et al.*, 2009). Historically, the application focused on the structure of matter. Recently, however, the application has been broadened, including other aspects of electromagnetic and non- electromagnetic radiation.

In the last twenty years, field spectrometry has been playing vital roles in characterizing the reflectance of vegetation types *in situ*, and providing a means of scaling up measurement at field (canopy and leaves) and laboratory levels (Milton *et al.*, 2009). Many attempts have been successfully made to discriminate and classify wetland species based on their fresh leaf reflectance at laboratory levels with the view to scaling it up to airborne remote sensing (Vaiphasa *et al.*, 2005; Vaiphasa *et al.*, 2007) and field reflectance at canopy scale (Best *et al.*, 1981; Penuelas *et al.*, 1993b; Schmidt and Skidmore, 2003; Becker *et al.*, 2005; Rosso *et al.*, 2005).

The earliest effort on spectral discrimination of wetland species was that of Anderson (1970) who attempted to evaluate the discrimination of ten marsh-plant species which dominated a wetland in Chesapeake Bay using ISCO Model SR Spectroradiometer. He concluded that the spectral difference between the species is minor in the visible spectrum, but significant in the near-infrared spectrum. The variation in the spectral reflectance with the changing seasons was also reported in the study. Best *et al.* (1981) investigated the use of four bands of Exotech radiometer to discriminate between the vegetation types which dominated the Prairie Pothole in the Dakotas. The spectral measurements were taken from ten common species during the periods of early-emergent, flowering, early-seed, and senescent phenological stages. Their findings showed that the best period to discriminate among the eight species studied was during the flowering and early-seed stages. However, it was difficult to differentiate reed (*Sparganium euryeapum)* from the other species. It was also concluded that a single species, in different phenological stages, showed significant variation in its spectral reflectance. Schmidt and Skidmore (2003) used the spectral reflectance measured at canopy level with A GER 3700 spectrometer from 27 wetland species to evaluate the potential of mapping coastal saltmarsh vegetation associations (mainly consisting of grass and herbaceous species) in the Dutch Waddenzee wetland. It was found that the reflectance in six bands distributed in the visible, near-infrared, and short wave infrared were the optimal bands for mapping saltmarsh vegetation (Table 2.2). Fyfe (2003) attempted to discriminate three coastal wetland species (*Zostera capricorni, Posidonia australis*, and *Halophila ovalis*) in Australia. Using a single-factor analysis of variance and multivariate techniques, it was possible to distinguish among the three species by their reflectance in the wavelengths between 530 nm–580 nm, 520 nm–530 nm, and 580 nm–600 nm. However, the differences were more significant between 570 and 590 nm. Rosso *et al.* (2005) in California, USA, collected spectral reflectance data from five species (*Salicornia, S. foliosa, S. alterniflora,* and *Scirpus)* using an Analytical Spectral Device (ASD) full-range (0.35nm –2.5 nm) PS II spectrometer to assess the separability of the marsh species under controlled conditions. Spectral Mixture Analysis (SMA) and Multiple Endmember

Spectral Mixture Analysis (MESMA) were used on the AVIRIS data. Using both SMA and MESMA, it was possible to distinguish between the species to achieve higher classification accuracies. However, the MESMA technique appeared to be more appropriate because it could incorporate more than one endmember per class. Similar work was also conducted by (Li *et al.*, 2005). They were able to use AVIRIS imagery to discriminate three salt marsh species (*Salicornia*, *Grindelia,* and *Spartina*) in China and in San Pablo Bay of California, USA. They developed a model that mixed the spectral angle together with physically meaningful fraction and the root mean square error. The results were satisfactory considering the success in discriminating the two marsh vegetation species (*Spartina* and *Salicornia)*, which covered 93.8% of the marsh area. However, it was difficult to discriminate *Grindelia* from *Spartina* and *Salicornia* due to the spectral similarity between the species. Becker *et al.* (2005) were able to use a modified version of the slope-based derivative analysis method to identify the optimal spectral bands for the differentiation of coastal wetland vegetation. They transformed hyperspectral data measured by the SE-590 spectroradiometer at canopy level into a second-derivative analysis. Six bands were found across the visible and near-infrared region to be powerful for discriminating the coastal wetland species.

In Thailand, Vaiphasa *et al.* (2005) were able to identify and distinguish 16 vegetation types in a mangrove wetland in Chumporn province. Their research was conducted by collecting hyperspectral reflectance data using a spectroradiometer (FieldSpec Pro FR, Analytical Spectral Device, Inc.), under laboratory conditions. The results of one-way ANOVA with a 95% confidence level ($p < 0.05$), and Jeffries–Matusita (JM) distance indicated that the best discrimination of the 16 species is possible with four bands located in the red-edge and near-infrared and mid-infrared regions of the electromagnetic spectrum (Table 2.2). Vaiphasa *et al.*, (2007) also used the same spectral data set to compare the performance of genetic algorithms (GA) and random selection using t-tests in selecting key wavelengths that are most sensitive in discriminating between the 16 species. The JM distance was used as an evaluation tool. The results showed that the separability of band combinations selected by GA was significantly higher than the class separability of randomly selected band combinations with a 95% level of confidence ($\alpha = 0.05$). Mangrove wetland species were also discriminated and mapped in Malaysia by Kamaruzaman and Kasawani (2007) who were able to use ASD Viewspec Pro-Analysis to collect the spectral reflectance data from five species at Kelantan and Terengganu, namely *Rhizophora apiculata*, *Bruguiera cylindrica, Avicennia alba, Heritiera littoralis,* and *Hibiscus tiliaceus.* The canonical stepwise discriminant analysis revealed that the five species were spectrally separable at five wavelengths (693 nm, 700 nm, 703 nm, 730 nm, and 731 nm) located in the red-edge and near-infrared region.

Wang *et al*. (2007) attempted to map highly mixed vegetation in salt marshes in the lagoon at Venice in Italy. Six significant bands of Compact Airborne Spectral Imager (CASI) were selected using Spectral Reconstruction (SR).The results showed that accuracy of Vegetation Community based Neural Network Classifier (VCNNC) can be used effectively in the situation of mixed pixels, thus, it yielded accuracy higher (91%) than the Neural Network Classifier (84%). Another attempt in discriminating marsh species was that by Artigas and Yang (2005) in the Meadowlands District in north-eastern New Jersey, USA. They conducted a study to characterize the plant vigour gradient using hyperspectral remote sensing with field-collected seasonal reflectance spectra of marsh species in a fragmented coastal wetland. Their results indicated that near-infrared and narrow wavelengths (670 nm-690 nm) in the visible region can be used to discriminate between most marsh species. However, it was difficult to discriminate between the two *Spartina* species because they belong to the same genus. It was concluded that these mixed pixels could be minimized using pixel unmixing techniques to discover the linear combinations of spectra associated with the pixels.

Table 2.2: Frequency of wavelengths selected in some studies for mapping wetland vegetation adapted into the four spectral domains defined by Kumar *et al*. (2001)

Wavelengths regions (nm)	Reference	Selected bands (nm)
Visible (400-700)	Daughtry and Walthall (1998)	550, 670
	Schmidt and Skidmore (2003)	404, 628
	Vaiphasa *et al.*(2005)	0
	Thenkabail *et al*. (2002)	490, 520, 550, 575, 660, 675
	Thenkabail *et al*. (2004)	495, 555, 655, 675
	Adam and Mutanga (2009)	0
Red-edge (680-750)	Daughtry and Walthall (1998)	720
	Schmidt and Skidmore (2003)	0
	Vaiphasa *et al.*(2005)	720
	Thenkabail *et al*. (2002)	700, 720
	Thenkabail *et al*. (2004)	705, 735
	Adam and Mutanga (2009)	745,746
Near-infrared (700-1300)	Daughtry and Walthall (1998)	800
	Schmidt and Skidmore (2003)	771
	Vaiphasa *et al.*(2005)	1277
	Thenkabail *et al*. (2002)	845, 905, 920, 975
	Thenkabail *et al*. (2004)	885,915,985,1085,1135, 1215,1245,1285
	Adam and Mutanga (2009)	892, 932, 934,958,961, 989
Mid-infrared (1300-2500)	Daughtry and Walthall (1998)	0
	Schmidt and Skidmore, (2003)	1398, 1803, 2183
	Vaiphasa *et al.*(2005)	1415, 1644
	Thenkabail *et al*. (2002)	0
	Thenkabail *et al*. (2004)	1445,1675, 1725, 2005, 2035, 2235, 2295, 2345
	Adam and Mutanga (2009)	0

In summary, most of the previous studies have stated that wetland vegetation has the greatest variation in the near-infrared and red-edge regions (Daughtry and Walthall, 1998; Cochrane, 2000; Schmidt and Skidmore, 2003; Vaiphasa *et al.*, 2005). Hence, most of the wavelengths selected to map wetland vegetation were mainly allocated in near-infrared and red-edge regions of the electromagnetic spectrum (Table 2.2).

More work is needed to build comprehensive spectral libraries for different wetland plants. Hyperspectral imagery proved to be useful in discriminating wetland species with higher accuracy. However, hyperspectral imagery is expensive to acquire, time-consuming to process, even when small areas are covered. Innovative new methods which take advantage of the relatively large coverage and high spatial resolution of the fine sensors and the high spectral resolution of hyperspectral sensors could result in more accurate discrimination models of wetland species at a reasonable cost.

2.7 Estimating biophysical and biochemical parameters of wetland species

The main biochemical constituents found in vegetation are nitrogen, plant pigment, and water. Whereas biophysical properties of the plant include LAI, canopy architecture and density, and biomass (Govender *et al.*, 2009), estimating the biochemical and biophysical properties of wetland vegetation is a critical factor for monitoring the dynamics of the vegetation productivity, vegetation stress, or nutrient cycles within wetland ecosystems (Asner, 1998; Mutanga and Skidmore, 2004a). The most important biochemical and biophysical properties that characterize the wetland species are: chlorophyll and biomass concentration, and leaf water content (Anderson, 1995). Few studies, however, have been conducted to study these properties that affect wetland plant canopies using both multispectral and hyperspectral remote sensing.

2.7.1 Mapping wetland biomass

Estimating wetland biomass is necessary for studying productivity, carbon cycles, and nutrient allocation (Zheng *et al.*, 2004; Mutanga and Skidmore, 2004a). Many studies of field biomass have used vegetation indices based on the ratio of broadband red and near-infrared reflectance. Ramsey and Jensen (1996) in the USA used a helicopter platform to measure spectra of the canopies of four species which dominated in south-west Florida to describe the spectral and structural change within and between the species and community types. Reflectance values were generated from the canopy spectral data to correspond with AVHRR (bands 1 and 2), Landsat TM (bands 1-4), and XMS SPOT (bands 1-3) sensors. The relationship between canopy structure and reflectance showed the difficulties of discrimination of mangrove species based on optical properties alone. Moreover, species composition was not correlated with any combination of reflectance bands or vegetation index. However, the study revealed the possibility of estimation of vegetation biomass such as LAI using red and near-infrared bands on various sensors.

Tan *et al.* (2003) used Landsat ETM bands 4, 3, and 2 false colour, and field biomass data to estimate wetland vegetation biomass in the Poyang natural wetland, China. Linear regression and statistical analyses were performed to determine the relationship among the field biomass data and some transformed data derived from the ETM data. Their results indicated that sampling biomass data has the best positive correlation to the Difference Vegetation Index (DVI) data. The authors developed a linear regression model to estimate the total biomass of the whole Poyang Lake natural conservation area. Similarly, Rendonga and Jiyuanb (2004) at Poyang in China, attempted to estimate the vegetation biomass in a large freshwater wetland using the combination of Landsat ETM data, GIS (for analyses and projecting both the sampling and Landsat ETM data), and GPS for (field biomass data). The results showed that the sampling of biomass data was best relative to the ETM 4 data with the highest coefficient of 0.86, at the significance level of 0.05. The study revealed that the near-infrared band could be used to estimate the wetland vegetation biomass.

The use of coarser spatial resolution sensors e.g. (VHR) IKONOS and AVHRR images has also been investigated in estimating wetland biomass. Proisy *et al.* (2007) created a new textural analysis method in which they applied Fourier-based Textural Ordination (FOTO) in 1 m panchromatic and 4 m infrared IKONOS images to estimate and map high biomass in forest wetland in French Guiana in the Amazon. Their work yielded accurate predictions of mangrove total aboveground biomass from both 1 m and 4 m IKONOS images. However, the best results were obtained from 1 m panchromatic with the maximum coefficient determination (R^2) above 0.87.

Moreau *et al.* (2003) investigated the potential and limits of two methods to estimate the biomass production of Andean wetland grasses in the Bolivian Northern Altiplano from NOAA/ AVHRR. The first method was based on monthly field biomass measurement and the second one was based on Bidirectional Reflectance Distribution Function (BRDF) normalized difference vegetation index (NDVI). Their results showed that BRDF normalized NDVI was sensitive to the green leaf or photosynthetically active biomass. The study also revealed that the optimal time for estimating the biomass with remotely sensed data in wetland species is during the growing season.

The limitations of using vegetation indices such as NDVI for estimation of biomass, especially where the soil is completely covered by the vegetation, have been reported in the literature. This is due mainly to the saturation problem (Thenkabail *et al.*, 2000; Mutanga and Skidmore, 2004a). Nevertheless, Mutanga and Skidmore (2004a) developed a new technique to resolve this saturation problem. They compared the use of band depth indices calculated from continuum-removed spectra with two narrow band NDVIs calculated using near-infrared and red bands to estimate *Cenchrus ciliaris* biomass in dense vegetation under laboratory conditions. The results clearly showed that band depth analysis approach proved to be efficient with a high coefficient in estimating biomass in densely vegetated areas where NDVI values are restricted by the saturation problem.

2.7.2 Estimation of leaf and canopy water content in wetland vegetation

Water availability is a critical factor in wetland plants' survival. There has been a rapid growth in remote sensing research to assess the vegetation water content as an indicator for the physiological status of plants, fire potential, and ecosystem dynamics at both laboratory and field levels using very high resolution spectrometers such as the ASD spectral device with spectral sampling intervals of less than 2 nm (Liu *et al.*, 2004; Stimson *et al.*, 2005; Toomey and Vierling, 2006). However, no significant research has been carried out on estimating water content in wetland plants especially. This is because the studies using remote sensing on wetland plants have been aimed mainly at discriminating and mapping, rather than estimating plant physiology such as water content and water stress.

Quite a number of different indices and techniques have been developed for estimating plant water content using the absorption features throughout the mid-infrared region (1300 nm-2500 nm) of the electromagnetic spectrum e.g. in the Netherlands (Zhang and Foody, 1998), Canada (Davidson *et al.*, 2006), and USA (Gao, 1996). The authors determined the canopy water content by scaling the foliar water content (FWC, %) with the specific leaf area (SLA), LAI, and the percent canopy cover for a specific forest canopy. However, Ceccato *et al.* (2001) noted that this technique relies on estimation of SLA, which varies according to species and phenological status.

Work by Penuelas *et al.* (1993a) found the water band index (WI), which has been developed based on the ratio between the water band 970 nm and reflectance at 900 nm, to be strongly correlated with relative plant water content. Using reflectance at 857 nm and 1241 nm, Gao (1996) developed the normalized difference water index (NDWI) in California, USA to estimate vegetation water. The results showed that the NDWI is less sensitive to atmospheric scattering effects than NDVI, and it is useful in predicting water stress in canopies and assessing plant productivity. It was recommended that further investigation was needed in order to understand this index better by testing it with the new generation of satellite instruments such as MODIS and SPOT-VEGETATION. Less sensitive semi-empirical indices for atmospheric scattering have also been developed by Datt (1999) to determine the relationship between spectral reflectance of several *Eucalyptus* species and both the gravimetric water content and equivalent water thickness (EWT). The results showed that EWT was significantly correlated with reflectance in several wavelength regions. However, no significant correlations could be obtained between reflectance and gravimetric water content.

The use of remote sensing in estimating plant water content is challenging because it is difficult to distinguish the contribution made by foliar liquid water and atmospheric vapour on the water-related absorption spectrum. This is because the absorption band related to water content is also affected by atmospheric vapour (Figure 1.1) (Liu *et al.*, 2004). Attempts have been made to minimize the atmospheric interference by using the red-edge position which is located outside the water absorption bands. In China, Liu *et al.* (2004) found a significant correlation between plant water content with the red-edge width in six different growth stages of wheat plants. The correlation coefficients were between 0.62 and 0.72 at 0.999 confidence level. The results were more reliable than those obtained using the WI and the NDWI. Similar results were reported in the USA by Stimson *et al.* (2005) who correlated foliar water content with the red-edge position to evaluate the relationship between foliar water content and spectral signals in two coniferous species: *Pinus edulis* and *Juniperus monosperma.* The results showed significant correlations of $R^2 = 0.45$ and $R^2 = 0.65$ respectively.

As there has been no significant research on estimating water content and water stress of wetland vegetation specifically, additional studies on these aspects are needed to better understand the spectral response of wetland plants. The results of such research could help the researcher to develop accurate models for describing, for example, the separability of wetland plant varieties and for estimating foliar nutrients and developing indicators that can quantify the integrated condition of wetland plants and can identify their primary stressors across a range of scales.

2.7.3 Estimating leaf area index of wetland vegetation

LAI is defined as the total one-sided area of all leaves in the canopy per unit ground surface area (m^2/m^2) (Gong *et al.*, 2003). Information on LAI is valuable for quantifying the energy and mass exchange characteristics of terrestrial ecosystems such as photosynthesis, respiration, evapotranspiration, primary productivity, and crop yield (Kumar *et al.*, 2001; Gong *et al.*, 2003). Research efforts on estimating LAI from spectral reflectance measurements have been focused mainly on forests (Gong *et al.*, 1995; Gong *et al.*, 2003; Pu *et al.*, 2005; Schlerf *et al.*, 2005; Davi *et al.*, 2006) and crops (Thenkabail *et al.*, 2000; Hansen and Schjoerring, 2003; Ray *et al.*, 2006). However, regardless of the work that has been done at Majella National Park, in Italy by, Darvishzadeh *et al.* (2008) the estimation of LAI for heterogeneous grass canopies has not been done. Moreover, a few studies dealing specifically with estimating LAI of wetland species have been conducted only in forest wetlands and mangrove wetlands (Green *et al.*, 1997; Kovacs *et al.*, 2004; Kovacs *et al.*, 2005).

In general, the above-mentioned studies have investigated several analytical techniques to estimate LAI using reflectance data. This can be grouped into two main techniques: the stochastic canopy radiation model and the empirical model. The empirical model has been more widely investigated than the stochastic canopy radiation model. The univariate regression analysis with vegetation indices such as NDVI and simple ratio, derived from visible and near-infrared wavelengths, is the most widely used empirical model and has been used in estimating LAI (Gong *et al.*, 1995; Green *et al.*, 1997; Thenkabail *et al.*, 2000; Gong *et al.*, 2003; Kovacs *et al.*, 2004; Kovacs *et al.*, 2005; Schlerf *et al.*, 2005).

Green *et al.* (1997) in UK developed a model based on gap-fraction analysis and NDVI derived from Landsat TM and SPOT XS to estimate LAI from three species: *Rhizophora mangle*, *Laguncularia racemosa*, and *Avicennia germinans* in a mangrove wetland in the West Indies. The model produced a thematic map of LAI with a high accuracy (88%) and low mean difference between predicted and measured LAI (13%).

Vegetation indices derived from high spatial resolution data were shown to be effective in monitoring LAI in mangrove forests. Kovacs *et al.* (2004) tested the relationship between *in situ* estimates of LAI and vegetation indices derived from IKONOS imagery in a degraded mangrove forest at Nayarit, Mexico. Regression analysis of the in situ estimates showed strong linear relationships between LAI and NDVI and simple ratio. Moreover, no significant differences were found between the simple ratio and NDVI models in estimating LAI at both plot sizes. In the same area, Kovacs *et al.* (2005) examined the potential of IKONOS in mapping mangrove LAI at the species level. A hand-held LAI-2000 sensor was also evaluated for the collection of data *in situ* on the mangrove LAI as a non-destructive alternative for the field data collection procedure. A strong significant relationship was found between NDVI, derived from IKONOS data, and *in situ* LAI collected with a LAI-2000 sensor. It was concluded that IKONOS satellite data and the LAI- 2000 could be an ideal method for mapping mangrove LAI at the species level.

Researchers have shown that vegetation indices (VIs) derived from the narrow band could be vital for providing additional information for quantifying the biophysical characteristics of vegetation such as LAI (Blackburn and Pitman, 1999; Mutanga and Skidmore, 2004a). In wetland environments specifically, however, only one work, that by Darvishzadeh *et al.* (2008) at Majella National Park in Italy, has investigated the use of hyperspectral data in estimating and predicting LAI for heterogeneous grass canopies. The study investigated the effects of dark and light soil and plant architecture on the retrieval of LAI red and near-infrared reflectance. Using A GER 3700 spectroradiometer, the spectral reflectances were measured from four different plant species (*Asplenium nidus, Halimium umbellatum, Schefflera arboricola Nora*, and *Chrysalidocarpus decipiens*) with different leaf shapes and sizes under laboratory conditions; then many VIs were calculated and tested. A stronger relationship was found between LAI and narrow band VIs in light soil than in dark soil. However, the narrow band simple ratio vegetation index (RVI) and second soil-adjusted vegetation index (SAVI2) were found to be the best overall choices in estimating LAI.

Although reasonable results were obtained from narrow band VIs in estimating LAI (Thenkabail *et al.*, 2000; Ray *et al.*, 2006; Darvishzadeh *et al.*, 2008), some authors noted that the strengths of a large number of hyperspectral bands have not yet been exploited by these methods because only two bands from red and near-infrared regions are used to formulate the indices (Hansen and Schjoerring, 2003; Schlerf *et al.*, 2005). A technique such as multiple linear regression (MLR) which uses the advantages of the high dimensionality of the hyperspectral data to select optimal band combinations to formulate VIs, was shown to be effective at estimating the biophysical and biochemical properties of vegetation such as LAI (Thenkabail *et al.*, 2000; Schlerf *et al.*, 2005).

Despite some success in estimating the biochemical and biophysical parameters in some ecosystems, estimation remains challenging in wetland environments where visible and near - infrared canopy reflectance has been revealed to be strongly affected by the background of soil and water, and atmospheric conditions. Further research is needed to develop indices that can reduce the effects of background and atmospheric quality.

2.8 Overall challenges and future research

Over the last few decades, considerable progress has been made in applying sensor techniques and data processing in discriminating, mapping, and monitoring wetland species. However, there are still challenges to be addressed in many aspects. First, traditional digital imagery from multispectral scanners is subject to limitations of spatial and spectral resolution compared to narrow vegetation units that characterize wetland ecosystems.

Second, despite the agreement on the effective performance of hyperspectral data in discriminating wetland species, the reflectances from different vegetation species are highly correlated because of their similar biochemical and biophysical properties. Furthermore, these properties are directly influenced by environmental factors and therefore the unique spectral signature of the plant species has become questionable (Price, 1994). In addition, spectral variations can also occur within a species because of age differences, micro-climate, soil and water background, precipitation, topography, and stresses.

Third, measurement of the biophysical and biochemical properties of vegetation using VIs derived from broad band sensors can be unstable due to the underlying soil types, canopy and leaf properties, and atmospheric conditions. For example, NDVI asymptotically saturate after a certain biomass density and for a certain range of LAI (Mutanga and Skidmore, 2004a). Hence, the measurement accuracy drops considerably (Gao *et al.*, 2000; Thenkabail *et al.*, 2000).

A fourth research challenge is that in most African countries (e.g. South Africa) there are only a handful of studies that have used hyperspectral data to characterize savanna vegetation due to high cost and poor accessibility (Mutanga *et al.*, 2003; Mutanga and Skidmore, 2004a; Mutanga and Kumar, 2007; Mutanga and Skidmore, 2007) Also, no research has yet been carried out on discriminating wetland vegetation and estimating its biophysical and biochemical parameters using process-based models that use remotely sensed data as input parameters.

Despite these shortcomings, there is no doubt that remote sensing technology could play a vital role in effectively discriminating and monitoring wetland species by selecting appropriate spatial and spectral resolution as well as suitable processing techniques for extracting spectral information of species.

From a research perspective, however, there are several major challenges in the application of remote sensing in wetland species that need to be addressed.

First, the most current remote sensing techniques in mapping vegetation have been undertaken in arid and semi-arid regions with low vegetation cover and less complexity within the vegetation unit. These techniques are, therefore, of little use for narrow vegetation units that characterize wetland ecosystems. Additional research effort is needed to adopt more classification techniques to improve the accuracy of the spatial resolution of the current sensors which varies from 20 m to 30 m. Hyperspectral radiometers are considered to be the sensors of choice in the future for mapping and monitoring wetland species. This has increased the need to build comprehensive spectral libraries for different wetland plant species under different plant conditions and environmental factors. Additionally, the fundamental understanding of the relationship between the reflectance measurements, wetland species' canopy density, and bottom reflectance parameters should be studied further. The spectral libraries of wetland species will help in discriminating not only between wetland species, but also between wetland species and upland species as there has been no specific research dealing with the difference in spectral response of canopies of wetland species and upland species.

Second, in the southern African region, more research is needed to enhance ability in discriminating wetland vegetation and estimating its biophysical and biochemical properties which have been overlooked in the scientific research. For example, papyrus swamps (*Cyperus papyrus L.*) (which characterize most of the tropical African wetlands, with a high rate of biomass production, a tremendous amount of combined nitrogen, that play vital roles in hosting habitats for wildlife and birds) are omitted in the application of remote sensing in discriminating wetland vegetation.

Third, although some studies have been undertaken on estimating the vegetation biophysical and biochemical parameters (e.g. LAI, water content, biomass, pigment concentration, and nitrogen) in different ecosystems, there is paucity of research on wetland species. After the progress in the field of spectrometry, researchers began to measure vegetation properties in complex ecosystems using new narrow band indices (Mutanga and Skidmore, 2004a) and red-edge position (Mutanga and Skidmore, 2007). These efforts should be further extended and developed so as to cope with wetland species environments where the saturation and the atmospheric vapour affect the near-infrared region. A fourth research prospect is the availability of hyperspectral sensors which could allow mapping both of species and their quality in wetland ecosystems. This will enhance a fundamental understanding of the spatial distribution of the quality and quantity of wetland species, which could lead to the development of early warning systems to detect any subtle changes in wetland systems such as signs of stress and lead to the development of techniques to classify wetland area conditions (e.g. healthy or disturbed) based on their species quality and quantity.

CHAPTER THREE

Spectral discrimination of papyrus (*Cyperus papyrus L.*) using a hand-held spectrometer under field conditions

This chapter is based on:

Adam, E., and Mutanga, O., 2009. Spectral discrimination of papyrus vegetation (*Cyperus papyrus L.*) in swamp wetland using field spectrometry. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64, 612-620.

Abstract

Techniques for mapping and monitoring wetland species are critical for their sustainable management. Papyrus (*Cyperus papyrus L.*) swamps are among the most important species rich habitats that characterize the Greater St Lucia Wetland Park (GSWP) in South Africa. This paper investigates whether papyrus can be discriminated from its co-existing species using ASD field spectrometer data ranging from 300 nm to 2500 nm, yielding a total of 2151 bands. Canopy spectral measurements from papyrus and other three species were collected *in situ* in the Greater St. Lucia Wetland Park, South Africa. A new hierarchical method based on three integrated analysis levels was proposed and implemented to spectrally discriminate papyrus from other species as well as to reduce and subsequently select optimal bands for the potential discrimination of papyrus. In the first level of the analysis using ANOVA, we found that there were statistically significant differences in spectral reflectance between papyrus and other species on 412 wavelengths located in different portions of the electromagnetic spectrum. Using the selected 412 bands, we further investigated the use of classification and regression trees (CART) in the second level of analysis to identify the most sensitive bands for spectral discrimination. This analysis yielded eight bands which are considered to be practical for upscaling to airborne or spaceborne sensors for mapping papyrus vegetation. The final sensitivity analysis level involved the application of the Jeffries–Matusita (JM) distance to assess the relative importance of the selected 8 bands in discriminating papyrus from other species. The results indicate that the best discrimination of papyrus from its co-existing species is possible with six bands located in the red-edge and near-infrared regions of the electromagnetic spectrum. Overall, the study concluded that spectral reflectance of papyrus and its co-existing species is statistically different, a promising result for the use of airborne and satellite sensors for mapping papyrus. The three step hierarchical approach employed in this study could systematically reduce the dimensionality of bands to manageable levels, a move towards operational implementation with band specific sensors.

Keywords: Papyrus. Greater St Lucia Wetland Park. Field spectrometer measurements. CART. Jeffries–Matusita.

3.1 Introduction

Papyrus (*Cyperus papyrus L.*) swamps characterizes most wetland areas of eastern and central tropical Africa (Bemigisha, 2004). Specifically, the swamp covers great areas in Uganda and Sudan around the Lake Victoria and Nile basins (Beadle, 1974). Other extensive areas are in the Upemba basin, Zaire, and the Okavango Delta, Botswana (Thompson *et al.*, 1979). Papyrus swamps usually create a buffer zone between terrestrial and aquatic ecosystems and play hydrological, ecological, and economic roles in the aquatic systems (Gaudet, 1980; Mafabi, 2000).

Previous studies found that tropical papyrus swamps are characterised by a tremendous amount of combined nitrogen (Muthuri and Kinyamario, 1989; Mwaura and Widdowson, 1992) and a high rate of biomass production (Muthuri and Kinyamario, 1989). In this regard, papyrus plays a vital role in hosting habitats for wildlife species such as the sitatunga antelope (*Tragelaphus spekei*) and African python (*Python sebae*) (Owino and Ryan, 2007). Papyrus also has some grazing potential and could be used as fodder with high nutritive value especially in the dry season when other forage is limited (Muthuri and Kinyamario, 1989). Further, studies found that the highest species richness of birds in marshland is associated with the areas where papyrus and natural vegetation were plentiful (Harper, 1992; Owino and Ryan, 2007). In addition to providing habitat for wildlife, the high biomass production characterizing papyrus swamps has been widely found to be useful for paper making. The Egyptians for example, were the first people who used papyrus to make paper more than five thousand years ago (Bucci, 2004). Recently, promising results have been obtained in using papyrus as an alternative source of fuel in many countries in central Africa such as Rwanda (Jones, 1983a; Muthuri and Kinyamario, 1989).

Despite its relative importance, human encroachment and intensified agricultural activities in many parts of Africa have threatened the existence of papyrus (Mafabi, 2000; Maclean *et al.*, 2006; Owino and Ryan, 2007). The continued degradation in papyrus habitat represents a significant threat to biodiversity conservation particularly for papyrus-specialist birds and other papyrus-reliant species in many African countries (Maclean *et al.*, 2006; Owino and Ryan, 2007).

To establish sustainable management of such important species, up-to-date spatial information about the magnitude and distribution of papyrus swamps at several scales is essentially required (Nagendra, 2001; Schmidt and Skidmore, 2003). This can be achieved through remote sensing techniques that can monitor the change in papyrus areas and assess the species' percentage covers as compared to the other species.

Traditionally, species discrimination for floristic mapping needs intensive fieldwork, including taxonomical information and the visual estimation of percentage cover for each species. This is costly and time-consuming and sometimes inapplicable due to the poor accessibility (Kent and Coker, 1994; Lee and Lunetta, 1995).   Remote sensing, on the other hand, is a technique that gathers the data regularly about the earth's features without actually being in direct contact with those features. The main advantages that make remote sensing preferable than field-based methods in land cover classification, is that it has repeat coverage which allows continuous monitoring, and its digital data can be easily integrated into Geographic Information System (GIS) for more analysis which is less costly and less time-consuming (Shaikh *et al.*, 2001; Ozesmi and Bauer, 2002; Schmidt and Skidmore, 2003; Mironga, 2004).

Both multispectral and hyperspectral remote sensing techniques have been used to discriminate and map wetland species. Multispectral data such as Landsat TM and SPOT imagery have been used to identify general vegetation classes or to attempt to discriminate just broad vegetation communities (May *et al.*, 1997; Harvey and Hill, 2001; Li *et al.*, 2005), while hyperspectral data have been successful in mapping wetlands vegetation at the species level (Schmidt and Skidmore, 2003; Brown, 2004; Rosso *et al.*, 2005; Belluco *et al.*, 2006; Kamaruzaman and Kasawani, 2007; Pengra *et al.*, 2007). Hyperspectral data have also been used to study vegetation health, water content in vegetation, biomass, and other physico-chemical properties (Green *et al.*, 1998; Ceccato *et al.*, 2001; Mutanga *et al.*, 2003; Mutanga and Skidmore, 2004a; Zarco-Tejada *et al.*, 2005).

In general, the use of multispectral data in discriminating and mapping wetlands species is challenging due to spectral overlap between the wetlands species and due to the lack of spectral and spatial resolution of the multispectral data (Rosso *et al.*, 2005). On the other hand, hyperspectral data often consist of over 100 contiguous bands of 10 nm or less bandwidth. These contiguous bands and narrow ranges lead to the possibility of discriminating and mapping vegetation species more accurately and precisely than the standard multispectral bands (Schmidt and Skidmore, 2003; Ustin *et al.*, 2004; Borges *et al.*, 2007).

A few previous attempts at using multispectral remote sensing in studies of papyrus swamps have been concentrated mainly on economic benefit and management scenarios of papyrus swamps, and promising results have been obtained (Bemigisha, 2004; Owino and Ryan, 2007). However, the spectral discrimination of papyrus (*Cyperus papyrus L.*) has been overlooked in scientific research. No attempt, to my knowledge, has been made to discriminate papyrus swamps using field spectrometry, let alone in South Africa where only a handful of studies have used hyperspectral data to characterize vegetation in general due to high cost and poor accessibility (Mutanga *et al.*, 2004; Ismail *et al.*, 2007).

Although hyperspectral data are critical in discriminating species, its high spectral resolution contains redundant information at band level (Kokaly *et al.*, 2003; Bajwa *et al.*, 2004). This high dimensional complexity of hyperspectral data can be problematic in terms of image processing algorithms, an excessive demand for sufficient field samples, high cost, and overfitting when using multivariate statistical techniques (Goetz, 1991; Bajcsy and Groves, 2004; Borges *et al.*, 2007; Mutanga and Kumar, 2007; Vaiphasa *et al.*, 2007). It is, therefore, imperative to identify the optimal bands required for discriminating and mapping wetland species without losing any important information. Different univariate and multivariate techniques for dimensionality reduction and band selection with different performance levels have been developed, such as canonical analysis, CART, discriminant analysis, principal component analysis, artificial neural network and Jeffries- Matusita (JM) (Satterwhite and Ponder Henley, 1987; Cochrane, 2000; Schmidt and Skidmore, 2003; Vaiphasa *et al.*, 2005; Milton *et al.*, 2009). However, inconsistent results have been obtained for different species and environments, and the use of a single technique in reducing data dimensionality to acceptable operational levels has not been very successful.

This study is aimed at investigating whether field spectrometry data could be used to effectively discriminate papyrus species from other species occurring in the swampy wetlands of South Africa. In other words, spectral separability analysis was used to examine whether papyrus swamps could spectrally be discriminated from each one of its co-existing species using field spectrometer measurements at canopy level as well as reducing spectral data dimensionality. More specifically, the objectives of this study were: 1. to determine whether there is a significant difference between the mean reflectance at each measured wavelength (from 350 nm to 2500 nm) for *Cyperus papyrus L.* and each one of the other co-existing three species (*Phragmites* *australis, Echinochloa pyramidalis,* and *Thelypteris interrupta*), and 2. To identify key wavelengths that are most sensitive in discriminating *Cyperus papyrus* from each one of the other three species. In order to achieve this, we used a field spectrometer to measure the spectral reflectance from papyrus swamps and the associated species in the Greater St Lucia Wetland Park in South Africa. To achieve an efficient optimal selection of bands, we propose a new hierarchical method that integrates Analysis of variance (first level), Classification regression trees (second level), and finally the Jeffries-Matusita distance analysis (third level) to assess the relative importance of the selected bands.

3.2 Material and methods

3.2.1 Field data collection

3.2.1.1 The identification of papyrus and its associated species

The most common plant species associated with papyrus in the swamps wetland in the study areas were identified in the field in the summer of 2007 under the supervision of an experienced ecologist using field observation techniques. These species were then recorded based on their density and estimation of percentage cover (covering at least 40 % of the area). In total, three species were identified as being the most co-existing species with papyrus. These were *Phragmites australis, Echinochloa pyramidalis,* and *Thelypteris interrupta* (Table 3.1).

Table 3.1: The papyrus swamp and its associated species, the number of sample plots and the total number of measurements collected

Species name	Type code	Nr of plots	Nr of measurements
Cyperus papyrus	CP	15	134
Phragmites australis	PA	9	111
Echinochloa pyramidalis	EP	7	101
Thelypteris interrupta	TI	10	113

3.2.1.2 Spectral data acquisition

The Analytical Spectral Devices (ASD) FieldSpec® 3 spectrometer was used to measure the spectral reflectance from papyrus and the other species. This spectrometer has a wavelength ranging from 350 nm to 2500 nm with a sampling interval of 1.4 nm for the spectral region 350 nm to1000 nm and 2.0 nm for the spectral region 1000 nm to 2500 nm, and a spectral resolution of 3 nm to 10 nm (ASD Analytical Spectral Devices Inc., 2005).

A combination of random sampling and purposive sampling was used to select field sites. Hawth's Analysis Tool extension for ArcMap designed to perform spatial analysis was used to generate random points in a land cover map developed using ASTER image. These points were then input in GPS to navigate to the field sites. Purposive sampling was done when the random point was not accessible, or to increase the variation of reflectance measurements of the species. Once the sampling location was indentified, a vegetation plot was defined to cover 3 m by 3 m in area of each species; then a total of 10 to 15 field spectrometer measurements were taken randomly from nadir at about 1.5 m and with a 5^{o} field of view above the vegetation species on each plot. This resulted in a ground field of view of about 13 cm in diameter, which was large enough to cover a cluster of species and to reduce the effects of background such as soil and water in the *in situ* spectral measurement (Table 3.1). All the measurements were collected in December 2007 between 10:00 am and 02:00 pm under sunny and cloudless conditions. A white reference spectralon calibration panel was used every 10 to 15 measurements to offset any change in the atmospheric condition and irradiance of the sun. Metadata such as the site description (coordinates and altitude, land cover class) and general weather conditions were also recorded to accompany field spectral measurements on each measured point (Milton *et al.*, 2009). Due to the atmospheric water absorption noise in the reflectance spectra, a number of bands around 1420 nm, 1940 nm, and 2400 nm were excluded from the analysis.

3.2.2 Data processing

It was difficult to use one technique to identify a reasonable number of wavelengths that are most sensitive from 350 nm to 2500 nm (n = 2151). This was because the dimensionality still remained high when one technique was used (412 wavelengths from analysis of variance). Moreover, there is no single technique that has universally proven to be superior for the optimal feature selection (Yang *et al.*, 2005), and it is quite possible that more than one subset of wavelengths can discriminate the data equally well (Yang *et al.*, 2005). We, therefore, innovated a new hierarchical method for spectral analysis based on three integrated levels.

3.3.2.1 First level (one-way ANOVA)

In the first level, we used one-way ANOVA to test if the differences in the mean reflectance between papyrus swamps and the other three species were statistically significant. We tested the research hypothesis that the means of the reflectance between the pairs of papyrus swamp and each one of the co-existing species (PA, EP, and TI) were significantly different at each measured wavelength, from 350 nm to 2500 nm, viz. the null hypothesis Ho: $\mu1 = \mu2$, $\mu1=\mu3$, $\mu1= \mu4$ versus the alternate hypothesis Ha: $\mu1 \neq \mu2$, $\mu1 \neq \mu3$, $\mu1\neq \mu4$ where: $\mu1$, is the mean reflectance values from papyrus and $\mu2$, $\mu3$, and $\mu4$ the mean reflectance values from *Phragmites australis, Echinochloa pyramidalis,* and *Thelypteris interrupta* respectively.

One- way ANOVA was used with a post-hoc Scheffé test at each measured wavelength for the individual class pair (CP *vs* PA, CP *vs* EP, and CP *vs* TI). We tested ANOVA with two confidence levels: a 99% confidence level ($p < 0.01$), and a 95% confidence level ($p<0.05$).

3.2.2.2 Classification and Regression Trees (CART)

We used CART in this second level of hierarchical methods to further reduce the number of the significant wavelengths obtained from ANOVA analysis, with the purpose of reducing data dimensionality. CART, which was developed by Breiman *et al.* (1984), is a non-parametric statistical model that can select from a large dataset of explanatory variables (**x**) those that are best for the response variables (**y**) (Yang *et al.*, 2003; Questier *et al.*, 2005). CART was preferred in this study because the values of the predictor variables (spectral reflectance) are continuous, as opposed to categorical target (plant species).

The CART model is built in accordance with the splitting rule. This rule performs the function of splitting the data into smaller parts according to the reduction of the deviance from the mean of the target variable (Y_{bar}) (or corrected total sum of the squares). (Y_i) is the target variable of each dataset. The decision tree begins a search from a root node (parent node) derived from all the predictors, and possible split points such that the reduction in deviance, *D* (total), is maximized (terminal node) as follows (Breiman *et al.*, 1984):

$$D\text{ (total)} = \sum(Y_i - Y_{\text{bar}})^2 \qquad (1)$$

The cut point, or value, always splits the data into two child nodes, the left node and the right node with maximum homogeneity. The reduction in deviance is as shown in the following equation:

$$\Delta_{\text{j,total}} = D\text{ (total)} - (D(\text{L}) + D(\text{R}) \qquad (2)$$

Where D(L) and D(R) are the deviances of the left and right nodes.

Hence, the algorithm begins searching for the maximized ($\Delta_{\text{j,total}}$) over all the predictor variables and the cut points subject to the constraint that the number of the members in the left and right nodes are larger than some criterion set by the user. The algorithm repeats the procedure of binary splitting for each node (left and right nodes) by treating each child node as a parent node splitting until the tree has a maximum size (Yang *et al.*, 2003).

In this study, we used CART as the second level of the hierarchical method to select the most sensitive wavelengths from the number of significant wavelengths selected in the first level (ANOVA). Therefore, CART generated the optimal bands by selecting only the spectral bands that result in small misclassification rates to discriminate each class pair (CP *vs* PA, CP *vs* EP, and CP *vs* TI) individually. The bands which were common in each class pair were then selected to get the optimal bands for all class pairs.

3.2.2.3 Distance analysis

After we had the optimal bands selected from the CART analysis, additional analysis was needed to identify the best band or band combinations that could be used for the best spectral separability between papyrus and each one of the three species. Hence, we tested the hypothesis that some bands are relatively more important than others in discriminating papyrus. The separability index used in this level of hierarchical method was the JM distance analysis (Schmidt and Skidmore, 2003; Ismail *et al.*, 2007; Vaiphasa *et al.*, 2007). It was impossible to run the JM distance analysis on all the significant bands (n = 412) from ANOVA analysis because of the singularity problem of matrix inversion (Vaiphasa *et al.*, 2005; Ismail *et al.*, 2007). Moreover, this high dimensional complexity is very costly, time-consuming, and beyond the capacity of the common image processing algorithms (Schmidt and Skidmore, 2003; Borges *et al.*, 2007; Vaiphasa *et al.*, 2007). We, therefore, used the bands derived from CART. The JM distance between a pair of probability functions is seen as quantification of the mean distance between the two class density functions (Richards and Jia, 2006). When classes are normally distributed, this distance turns out to be the Bhattacharyya (BH) distance (Schmidt and Skidmore, 2003; Richards and Jia, 2006). The JM distance has upper and lower bounds that vary between 0 and $\sqrt{2}$ (≈ 1.414), with the higher values indicating the total separability of the class pairs in the bands being used (ERDAS, 2005; Richards and Jia, 2006). In this study, we decided to use higher separability values ≥ 97 % as the JM distance threshold to identify the most important band or band combinations for best discrimination of papyrus swamps. The formula for computing the JM distance is as follows (ERDAS, 2005):

$$\alpha = \frac{1}{8}\left(\mu_i - \mu_j\right)^T \left(\frac{C_i + C_j}{2}\right)^{-1} \left(\mu_i - \mu_j\right) + \frac{1}{2}\ln\left(\frac{\left(\left|C_i + C_j / 2\right|\right)}{\sqrt{\left|C_i\right| * \left|C_j\right|}}\right) \qquad (3)$$

$$JM_{ij} = \sqrt{2\left(1 - e^{\alpha}\right)} \qquad (4)$$

Where:

i and j = the two classes being compared

C_i = the covariance matrix of signature i

μ_i = the mean vector of signature i

Ln = the natural logarithm function

$\left|C_i\right|$ = the determinant of C_i (matrix algebra).

3.3 Results

3.3.1 First level: ANOVA test

ANOVA results indicate that there is no significant difference between the two class pairs (CP *vs* EP, and CP *vs* TI) when a 99% confidence level (p< 0.01) was used. However, the 95% confidence level (p < 0.05) indicated that there is a statistically significant difference in the spectral reflectance between all the class pairs (CP *vs* PA, CP *vs* EP, and CP *vs* TI) at n = 412 wavelengths. These significant wavelengths were highlighted using a histogram for every individual class pair. The results of ANOVA test for each class pair (CP *vs* PA, CP *vs* EP, and CP *vs* TI) are shown in Figure 3.1 (a, b, and c). The shaded areas show the wavelengths where the spectral reflectance from the papyrus swamp is statistically different from the other three species, with a 95% confidence level (p-value < 0.05).

The conclusions from the ANOVA test are that the mean reflectance between papyrus and the other three species is significantly different in many measured wavelengths. These significant wavelengths are located in three different regions of the electromagnetic spectrum (red- edge, near-infrared, and mid-infrared).

Table 3.2 shows the frequency of the significant bands adapted into the four spectral domains which is widely used in the hyperspectral remote sensing of vegetation (Kumar *et al.*, 2001). The table shows that there are no statistically significant wavelengths located in the visible region for the class pairs CP *vs* EP, and CP *vs* TI. However, the class pair CP *vs* PA has more significant wavelengths located all over the spectral regions than any other class pair (CP *vs* EP, and CP *vs* TI). All the wavelengths from 350 to 1300 (n = 950) are significant for CP *vs* PA as well as 49.95% (n = 600) of wavelengths located in the mid-infrared region, whereas the statistically significant wavelengths for the pair CP *vs* TI are located only in the red-edge and near-infrared portions of the electromagnetic spectrum (n = 449).

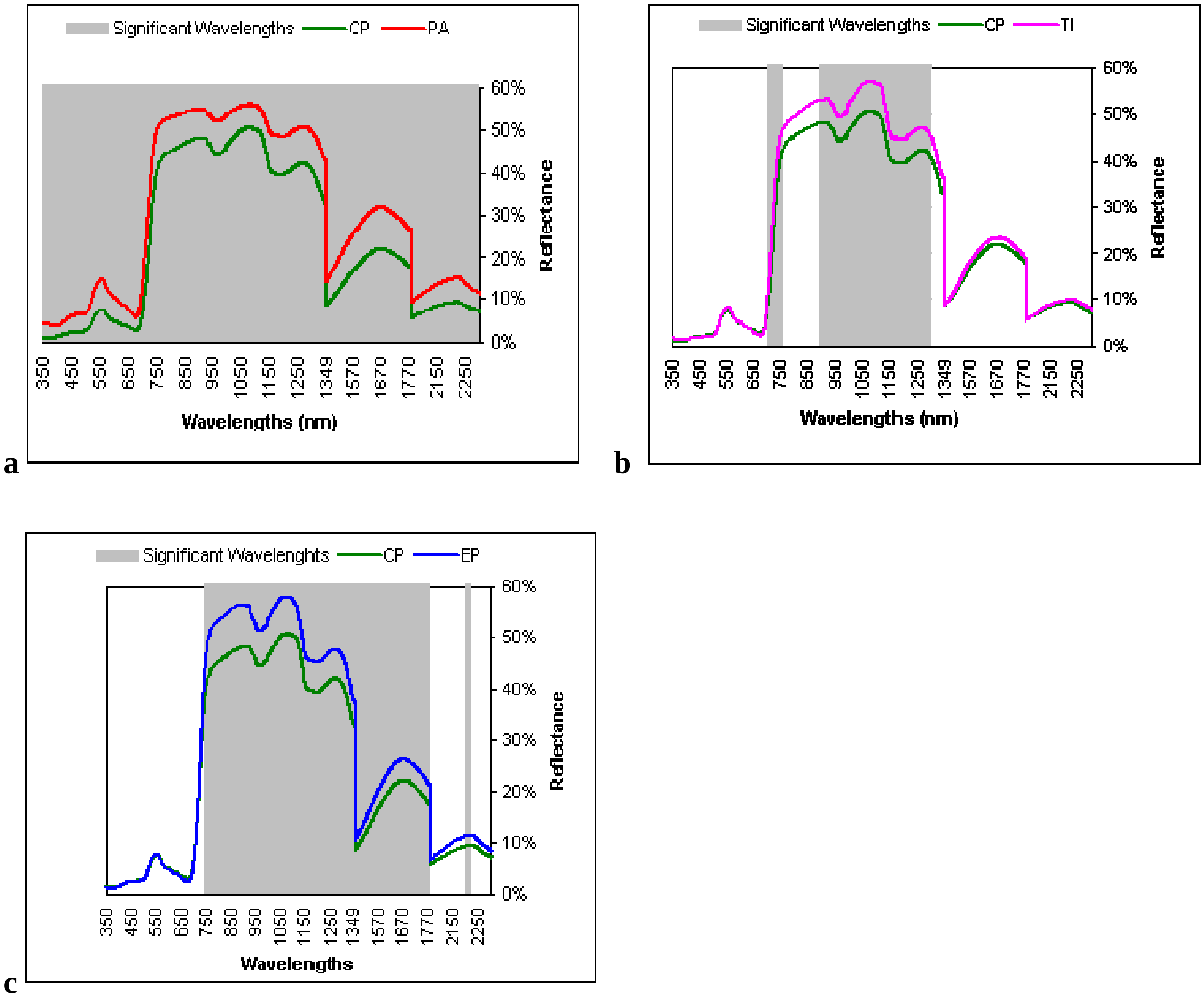

Figure 3.1. ANOVA results for each class pair (a) CP vs. PA, (b) CP vs. EP, and (c) CP vs. TI. The grey areas show the wavebands where there are significant differences between the class pairs within the electromagnetic spectrum.

Table 3.2: Frequency of significant bands for each class pair adapted into the four spectral domains defined by Kumar *et al*. (2001)

Wavelength region (nm)	Description	Band No	Significant bands					
			CP *vs* PA	%	CP *vs* EP	%	CP *vs* TI	%
350-700	Visible	351	351	100.00	0.00	0.00	0.00	0.00
680-750	Red-edge	71	71	100.00	10	14.08	45	63.38
700-1300	Near-infrared	601	601	100.00	560	93.18	451	74. 04
1300-2500	Mid-infrared	1201	600	49.95	367	30.55	0.00	0.00

It can also be seen from Table 3.2 that the red-edge and near-infrared are the most important regions where each class pair has the most statistically significant wavelengths. The results can be clearly seen in the histogram in Figure 3.2 which shows by maximum grey shading the wavelengths with the maximum frequency. These significant wavelengths have the potential to discriminate papyrus species from all other species (PA, EP, and TI).

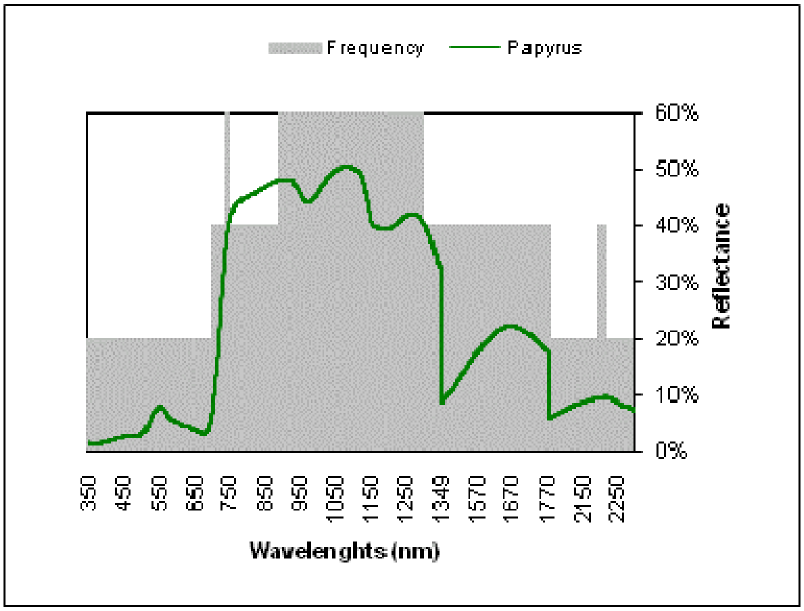

Figure 3.2. Frequency of statistical difference using ANOVA with 95% confidence level (P < 0:05) between the mean reflectance of papyrus and all other species. The maximum grey shading shows the wavelengths where papyrus could be discriminated from all the other three species.

Results of frequency analysis (Figure 3.2) reveal that there is no wavelength that maximized the discrimination of papyrus from the other species in the visible region. There are however, a few significant wavelengths located in the red-edge (741-746) nm = (n = 6) and a majority of wavelengths located in the near-infrared part of the electromagnetic spectrum (982-1297) nm = (n = 406). Further analysis was then conducted to reduce the number of these significant wavelengths (n = 412).

3.3.2 Second level: CART results

CART analysis was applied to reduce the numbers of significant bands (n = 412) selected by ANOVA analysis to fewer bands that could optimally discriminate the papyrus from the other three species. The selection of the optimal wavelengths was done for each individual class pair: CP *vs* PA (n = 17), CP *vs* EP (n = 13), and CP *vs* TI (n = 15). The misclassification rate was 0.014, 0.014, and 0.029 for each class respectively. The results are shown in Table 3.3.

Table 3.3: Wavelengths selected by CART for each individual class pair and the misclassification rate. Wavelengths that were able to differentiate between all three pairs of classes are highlighted in grey

Class pair	Wavelengths (nm) selected	No of wavelengths (nm)	Misclassification rate
CP *vs* PA	741, 745, 746, 892, 932, 934, 958, 961, 985 989, 1037, 1107, 1120, 1125, 1130, 1153, 1291.	17	0.014
CP *vs* EP	745, 746, 892, 932, 934, 958, 961, 989, 1056, 1119, 1123, 1124, 1153.	13	0.014
CP *vs* TI	741, 745, 746, 892, 932, 934, 958, 961, 989, 1010, 1038, 1056, 1119, 1130, 1146.	15	0.029

The common wavelengths among all class pairs (CP *vs* PA, CP *vs* EP, and CP *vs* TI) were then selected to find the optimal wavelengths for all class pairs. It also interesting to note that in Table 3.3 there are eight spectral bands that appeared commonly in every class pair. These spectral bands are: 745 nm, 746 nm, 892 nm, 932 nm, 934 nm, 958 nm, 961 nm, and 989 nm. From this analysis, these eight wavelengths could potentially discriminate papyrus species from all the three species.

3.3.3 Third level: Distance analysis results

The Table 3.4 shows the results of the JM distance analysis. The band located at 892 nm appeared to be the best single band because it produces best separability when used individually with a JM value of 1.342 (94.91%). Furthermore, it has the highest frequency (100 %) by appearing in every best band combination. The table also reveals that the use of more bands improves the separability of the papyrus. Whereas the single band (892 nm) produces an unacceptable JM value of 1.342, the acceptable average JM values (≥ 97 %) are reached when using three band combinations, which achieved 97.45%. The JM value then improved considerably until it reached the best value with the best eight band combinations.

Table 3.4: The averages of JM distance analysis for all the three class pair (CP vs PA, CP vs EP, and CP vs TI). The symbol (X) indicates the selection of optimal bands in each band combination

Best band combinations	745	746	892	932	934	958	961	989	JM value	%
Single band			X						1.342	94.91
Two bands			X		X				1.362	96.32
Three bands			X		X			X	1.378	97.45
Four bands			X		X	X	X		1.386	98.01
Five bands	X	X	X			X	X		1.393	98.51
Six bands	X	X	X		X	X	X		1.402	99.15
Seven bands	X	X	X	X		X	X	X	1.409	99.65
Eight bands	X	X	X	X	X	X	X	X	1.411	99.79

Table 3.5 shows the JM distance values for each individual class pair (CP *vs* PA, CP *vs* EP, and CP *vs* TI) within each best band combination. For the class pairs, CP *vs* PA and CP *vs* TI, a single band located at 892 nm produced an acceptable JM distance value. However, the class pair, CP *vs* EP, reached the acceptable value of JM distance (≥ 97 %) only when using six band combinations located at 745 nm, 746 nm, 892 nm, 934nm, 958 nm, and 961nm, where the other two class pairs (CP *vs* PA and CP *vs* TI) reached total separability of 100% (upper JM value). Unlike the other two class pairs (CP *vs* PA and CP *vs* TI), the CP *vs* EP pair does not reach the total separability even when using all the eight bands (JM distance value 1.405). However, total separability starts for the other two class pairs (CP *vs* PA and CP *vs* TI) from using the best four band combinations located at 892 nm, 934nm, 958 nm, and 961nm.

Table 3.5: The values of the JM distance for each individual class pair within the selected best band combinations.

Best combination	CP vs PA		CP vs EP		CP vs TI	
	JM value	%	JM value	%	JM value	%
892.	1.409	99.64	1.210	85.57	1.408	99.58
892, 934.	1.412	99.86	1.263	89.32	1.410	99.72
892, 934, 898.	1.413	99.93	1.308	92.50	1.413	99.93
892,934, 958, 961.	1.414	100.00	1.329	93.99	1.414	100.00
745, 745, 892, 958, 961.	1.414	100.00	1.351	95.55	1.414	100.00
745,745, 892, 934, 958, 961.	1.414	100.00	1.379	97.52	1.414	100.00
745, 746, 892, 932, 958, 961, 989.	1.414	100.00	1.399	98.94	1.414	100.00
745, 746, 892, 932, 934, 958, 961, 989.	1.414	100.00	1.405	99.36	1.414	100.00

3.4 Discussion

The use of field spectrometry for species discrimination is widespread at both field measurement and laboratory levels (Skidmore *et al.*, 1988; Schmidt and Skidmore, 2003; Brown, 2004; Rosso *et al.*, 2005; Vaiphasa *et al.*, 2005; Belluco *et al.*, 2006; Pengra *et al.*, 2007). The removal of redundant data and identification of relevant data are critical considerations in field spectrometry data processing. One should seek to ensure that this dimensionality reduction would not cause any loss of important information relevant to the object under study. Various researchers have used different techniques with inconsistent results to identify important bands of the electromagnetic spectrum for discriminating vegetation species.

In this paper, it was difficult to use one technique to identify a reasonable number of wavelengths that are most sensitive from 2150 bands, because the dimensionality remained still high when only one technique was used (412 wavelengths from analysis of variance). This could be explained by, firstly, the agreement that there is no single technique that has universally proven superior for the optimal feature selection (Yang *et al.*, 2005) and, secondly, the possible existence of a different subset of features that discriminates the data equally well (Yang *et al.*, 2005). Hence, a new hierarchical method was developed based on the integration of three analysis levels (ANOVA, CART, and JM) to reduce the dimensionality in the collected field spectrometry measurement data to discriminate papyrus from three other species. This is an important prerequisite for mapping papyrus swamps using airborne and satellite hyperspectral sensors. Results of this study show that the discrimination of papyrus from its associated species is possible at the field level using field spectrometry.

3.4.1 Differences in mean reflectance between papyrus and its associated species

The results from ANOVA test presented in Figure 3.1 and Table 3.2 have shown that there is a significant difference in the mean reflectance between papyrus and each of the three species studied (PA, EP, and TI) in the red-edge, near-infrared, and mid-infrared regions. The wavelength regions with the greatest frequency of significant differences between papyrus and other species can be seen in a histogram in Figure 3.2. These significant wavelengths are located in the red-edge region from 741nm to746 nm (n = 6) and in the near-infrared region from 892 to1297 nm (n = 406). This confirms the results of previous studies that state that green leaves have greatest variation in the near-infrared and red-edge regions (Asner, 1998; Daughtry and Walthall, 1998; Cochrane, 2000; Schmidt and Skidmore, 2003; Thenkabail *et al.*, 2004; Vaiphasa *et al.*, 2005). Although no leaf biochemical properties were directly measured in this study, it is likely that the occurrence of significant wavelengths in the Red-edge region (680 nm to 750 nm) is due to the variation between papyrus and other species on chlorophyll concentration, nitrogen concentration, and water content (Curran *et al.*, 1990; Curran *et al.*, 1991; Filella and Penuelas, 1994; Mutanga and Skidmore, 2007). This is because of the physiological evidence that papyrus is characterized by a tremendous amount of combined nitrogen, higher chlorophyll concentration, and higher rates in biomass production than most other wetlands species (Muthuri and Kinyamario, 1989; Mwaura and Widdowson, 1992). Unlike other species, papyrus is basically restricted to the area that is permanently either wet or flooded throughout the year. This results in a higher water content in a papyrus leaf compared to the other species. It is, therefore, assumed that the chlorophyll and nitrogen concentrations and water content vary significantly between papyrus and other species. The significant wavelengths in the near-infrared region, on the other hand, may be due to variation between papyrus and other species in the canopy structure (Kumar *et al.*, 2001; Schmidt and Skidmore, 2003). The differences in canopy and leaf structure of the different species are shown in Figure 3.3.

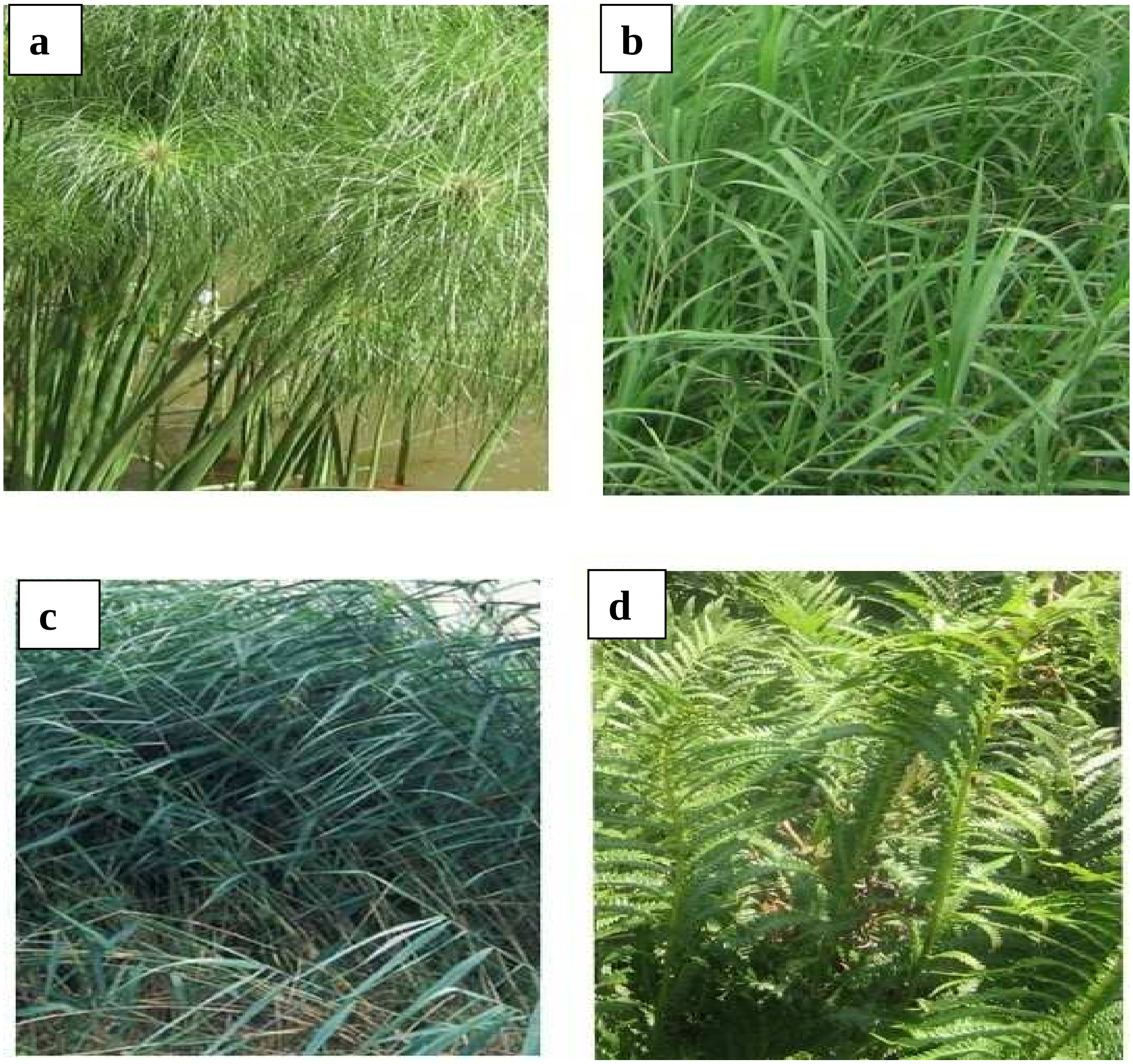

Figure 3.3. Variations in canopy and leaf structure in the four species: (a) *Cyperus papyrus,* (b) *Echinochloa pyramidalis,* (c) *Phragmites australis,* and (d) *Thelypteris interrupta*. Surface leaf structure in *Cyperus papyrus* is relatively most different from the other species.

3.4.2 Band selection using classification and regression trees (CART)

CART has helped to reduce dimensionality in the significant wavelengths (n = 412) obtained from ANOVA as well as to identify the most sensitive wavelengths to discriminate papyrus (Breiman *et al.*, 1984; De'ath and Fabricius, 2000; Questier *et al.*, 2005; van Aardt and Norris-

Rogers, 2008). As we aimed to discriminate only papyrus, CART was applied for each class pair individually (CP *vs* PA, CP *vs* EP, and CP *vs* TI). Table 3.3 shows the bands selected and the misclassification error rate. Relative to other studies, the misclassification error rate of this study is very low (De'ath and Fabricius, 2000; Questier *et al.*, 2005; van Aardt and Norris-Rogers, 2008). Therefore, we conclude that the selected bands in this analysis level are optimal bands for discriminating papyrus. The selected wavelengths were compared to wavelengths selected in the other previous studies as shown in Table 2.2. From Table 2.2 one can note that the bands selected not only in this study but also in the previous studies do not totally coincide with one another. This is explained mainly by the variation in concentration of pigments and the other optical properties and biochemical contents of the leaves between species, which leads to the different interactions within wavelengths of the electromagnetic regions (Asner, 1998; Kumar *et al.*, 2001; Schmidt and Skidmore, 2003) However, general trend, especially within the red-edge and near-infrared regions, does exist between the studies which reveal the relative importance of using different wavelengths of electromagnetic spectrum for species discrimination.

The study also confirms the advantages of CART (De'ath and Fabricius, 2000). This is can be summarized as being : 1. a simple, easy, and fast nonparametric method regarding the input data and output, 2. in variance to monotonic transformation of the explanatory variables, and 3. flexible in handling different dependent variables and highly discriminatory data. This data can be easily separated into individual classes or ignored without influencing the predication.

3.4.3 The JM distance analysis

The JM distance analysis was used to assess the relative importance of band combinations in discriminating between papyrus and other species (CP vs PA, CP vs EP, and CP vs TI) using bands selected by CART. We opted to use higher acceptable separability values (≥ 97 %) rather than ≥ 95 % (Vaiphasa *et al.*, 2005). This was done in order to achieve a precise selection of the most sensitive bands to discriminate papyrus. We found that some bands have more power for discriminating between papyrus and the other three species by having higher values of the JM distance. This is clearly shown in Table 3.4, which shows that three bands located at 892 nm, 934 nm, and 989nm can produce acceptable average separability (97.45%). The two class pairs (CP *vs* PA and CP *vs* TI) are spectrally more distant than the other class pair (CP *vs* EP) as is shown in Table 3.5. Papyrus, therefore, has greater potential of being separable from these two species (PA and TI) even with a single band located at 892nm.This is explained by the differences in the distance separability between the vegetation species (Skidmore *et al.*, 1988). As shown in Table 3.5, increasing the number of bands leads to an increase in the distance between the class pairs. For example, the four bands located at 892nm, 934nm, 958nm, and 961nm show maximum JM values for the two class pairs, CP *vs* PA and CP *vs* TI. These maximum values (as measured using the JM distance) indicate best discrimination between papyrus and the two species at these selected bands. CP and EP are similar in spectra. Therefore, only six band combinations located at 961nm, 745 nm, 934 nm, 746 nm, 892 nm, and 958 nm have the acceptable separability for the class pair, CP *vs* EP. These six bands have the potential to discriminate papyrus from all its co-existing species. These numbers of bands are consistent with previous studies that state that the best six band combinations have the greatest potential for better species discrimination (Schmidt and Skidmore, 2003). The results from this distance analysis predict the potential of correct discrimination of papyrus from its co-existing species using hyperspectral remote sensing (Schmidt and Skidmore, 2003; Vaiphasa *et al.*, 2005).

3.5 Conclusions

From this study we can conclude that:

1. Field spectrometer measurements at canopy level can be used to discriminate *Cyperus papyrus L.* from *Phragmites australis*, *Echinochloa pyramidalis,* and *Thelypteris interrupta.* This implies that the mean spectral reflectance of *Cyperus papyrus* is different from the other species associated with it in the same ecosystem (swamp wetlands).
2. CART can be used to considerably reduce the dimensionality and to select the most important bands for discriminating papyrus from the other species with a low rate of misclassification.
3. The use of CART has revealed that the greatest discrimination power for papyrus is located in the red-edge and near-infrared regions, specifically at 745 nm, 746 nm, 892 nm, 932 nm, 934 nm, 958nm, 961nm, and 989nm. This shows the importance of the red-edge and near-infrared regions in species discrimination, thereby confirming previous studies that found strong spectral variation among the vegetation species in these regions of the electromagnetic spectrum.
4. Although a single band located at 892 nm can discriminate *Cyperus papyrus* from *Phragmites australis* and *Thelypteris interrupta*, only six bands located at 745 nm, 746 nm, 892 nm, 934 nm,

958nm, and 961nm, show the potential to discriminate *Cyperus papyrus* from *Echinochloa pyramidalis*.

Overall, results of this study offer the possibility of extending field measurements at canopy level to airborne and satellite hyperspectral sensors data for discriminating *Cyperus papyrus* in swamp wetlands in South Africa. Further studies are also necessary to investigate the use of more advanced models such as the RF algorithm to discriminate among papyrus and its co-existing species (multi-class classification).

Acknowledgement

We would like to thank the Herbarium at University of KwaZulu-Natal for the support in the field planning and the identification of plant species. We extend our gratitude to Riky Taylor and Nerosha Covender in GSLWP for providing excellent conditions for the success of the field data collection. Our thanks go to Craig Morris and Elfatih Abdel-Rahaman for facilitating the task.

CHAPTER FOUR

Spectral discrimination of papyrus (*Cyperus papyrus L.*) and its co-existing species using hyperspectral data

This chapter is based on:

1. **Adam,** E., Mutanga, O., Rugege, D., and Ismail, R., 2009. Field spectrometry of papyrus vegetation (*Cyperus papyrus L.*) in swamp wetlands of St. Lucia, South Africa. *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, IV-260 – IV-263.
2. **Adam,** E., Mutanga, O., Rugege, D., and Ismail, R., 2010. Discriminating the papyrus vegetation (*Cyperus papyrus L.*) and its co-existent species using RF and hyperspectral data resampled to HYMAP. *International Journal of Remote Sensing*, (in press).

Abstract

Techniques for discriminating swamp wetland species are critical for the rapid assessment and proactive management of wetlands. In this study, we tested whether the RF algorithm could discriminate between papyrus vegetation and its co-existent species (*Phragmites australis*, *Echinochloa pyramidalis,* and *Thelypteris interrupta*) using *in situ* canopy reflectance spectra. Canopy spectral measurements were taken from the species using Analytical Spectral Devices but later resampled to Hyperspectral Mapper (HYMAP) resolution. The RF algorithm and a simple forward variable selection technique were used to identify key wavelengths for discriminating papyrus swamp and its co-existing species. The method yielded ten wavelengths located in the visible and SWIR portions of the electromagnetic spectrum with lowest out-of-bag estimate error rate of 9.5 % and .632+ bootstrap error of 8.95 %. The use of RF as a classification algorithm resulted in overall accuracy of 90.5 % and a KHAT value of 0.87, with individual class accuracies ranging from 93. 73 % to 100 %. Additionally, the results from this study indicate that the RF algorithm produces better classification results than conventional classification trees when using all HYMAP wavelengths (n = 126) and when using wavelengths selected by the forward variable selection technique.

Keywords*:* Papyrus swamp. Random forest. Classification trees, HYMAP sensor. Mapping, Variable selection. .632+ bootstrap.

4.1 Introduction

Wetland vegetation is an important component of wetland ecosystems that play hydrological, ecological, and economic roles in aquatic systems (Kokaly *et al.*, 2003; Yuan and Zhang, 2006). Wetland vegetation is an excellent indicator for early signs of any physical, chemical, and biological degradation in wetland environments (Dennison *et al.*, 1993; Zomer *et al.*, 2009). Furthermore, the distribution of wetland vegetation is an important factor influencing the feeding patterns and the distribution of wildlife in a wetland ecosystem. For example, in the Greater St Lucia Wetland Park, South Africa, papyrus (*Cyperus papyrus L.*) swamp forms critical habitats for a large number of species and several communities such as the Common Hippopotamus (*Hippopotamus amphibious)*, Nile crocodile (*Crocodylus niloticus)*, Great White Pelican (*Pelecanus onocrotalus*), and Pink-backed pelican (*Pelicanus rufescen*) (Grenfell *et al.*, 2009). Researchers have also noted that papyrus swamps play a vital role in intercepting the materials moving from catchments to open water (Azza *et al.*, 2000; Serag, 2003; Kyambadde *et al.*, 2004). Moreover, promising results have been obtained in using wetland species such as papyrus as an alternative source of fuel in many countries in central Africa, such as Rwanda (Jones, 1983a; Muthuri and Kinyamario, 1989).

Despite the remarkably rich biodiversity of papyrus swamps, their conservation and protection are a neglected issue in Africa (Owino and Ryan, 2007). As a result, human encroachment and intensified agricultural activities in many parts of Africa have threatened the existence of papyrus (Mafabi, 2000; Maclean *et al.*, 2006; Owino and Ryan, 2007). Effective techniques for mapping and monitoring papyrus swamp and its co-existing vegetation species are therefore critical for a better understanding of the magnitude and the distribution of papyrus and its co-existing species at several scales (Pengra *et al.*, 2007). However, wetland areas are generally difficult to map and to monitor due to poor accessibility, and sometimes they host both dangerous wildlife and endemic diseases (Zomer *et al.*, 2009). Additionally, traditional methods available for mapping plant species require intensive fieldwork and laboratory analysis for measuring the biochemical and biophysical properties of vegetation species (Mutanga *et al.*, 2003).

This usually results in the collection and analysis of data that are not generally representative of the plant population, especially if large and highly diverse areas such as wetlands are investigated (Mutanga *et al.*, 2003; Lawrence *et al.*, 2006).

Remote sensing potentially offers an economical and alternative method of discriminating amongst papyrus swamp and its co-existent vegetation species by reducing the intensive field sampling and laboratory analysis required by traditional mapping techniques. However, identifying wetland plant species is challenging using multispectral imagery due to the lack of spatial resolution of most of the current satellites with respect to the small and sharp vegetation units present in wetland environments (May *et al.*, 1997; Harvey and Hill, 2001; McCarthy *et al.*, 2005). Hence, with multispectral imagery the majority of pixels are mixtures of several vegetation species in various proportions (Zomer *et al.*, 2009). Additionally, the use of broad spectral bands of multispectral imagery for mapping wetland species remains difficult due to the spectral overlap between the species, because healthy vegetation species generally exhibit similar spectral responses in the visible and near-infrared region due to similar and limited basic components that contribute to their spectral reflectance (Price, 1992; Kokaly *et al.*, 2003). Furthermore, the canopy reflectance spectra of wetland vegetation are combined with reflectance spectra of the underlying soil and hydrologic regime (Yuan and Zhang, 2006). This combination usually results in a decrease in the spectral reflectance, especially in the near-to mid-infrared regions where water absorption is stronger (Fyfe, 2003; Silva *et al.*, 2008). Recent advances in airborne imaging sensors, in particular high spectral resolution hyperspectral platforms, such as HYMAP, offer the potential to discriminate wetland vegetation at species level due to the availability of narrow spectral channels of less than 10 nm (Schmidt and Skidmore, 2003; Rosso *et al.*, 2005; Vaiphasa *et al.*, 2005; Vaiphasa *et al.*, 2007). These narrow spectral channels permit an in-depth detection of detailed vegetation species which are otherwise masked by the broad wavebands acquired using multispectral data (Mutanga *et al.*, 2003; Vaiphasa *et al.*, 2005). However, the advantages of utilizing hyperspectral data also come with challenges in data processing and analysis which may lead to poor performance or even failure of the classification algorithm (Kavzoglu and Mather, 2002; Tsai *et al.*, 2007).

The discriminating of papyrus from each one of its co-existing species (binary class classification) at canopy level has been achieved using spectrometer measurements with a spectral sampling interval of less than 2 nm (Adam and Mutanga, 2009). Spectrometry (also known as spectroscopy) data is mostly acquired using hand-held, airborne, and spaceborne sensors. A hand-held spectroradiometer is an optical instrument used for measuring the spectrum emanating from a target in one or more fixed wavelengths in the laboratory and in the field

(Mutanga, 2005). Nevertheless, the current operational airborne and spaceborne sensors, such as HYMAP, lack fine spectral resolution (Mutanga, 2005). Therefore, it might be useful if the potential of specific spectral bands of these sensors in discriminating among papyrus and it co-existing species (Multi class classification) are investigated, through resampling fine spectral resolution data from spectrometers to coarser spectral resolutions of the spaceborne sensors. If the results are positive, the mapping and monitoring of wetland plant species could be operational on satellite platforms.

One of the most notable difficulties in hyperspectral data processing is the increase of data dimensionality, which requires sufficient training samples (Borges *et al.*, 2007; Hsu, 2007; Tsai *et al.*, 2007). Practically, in most of the hyperspectral applications the number of training samples (*n*) is limited with respect to the large number of hyperspectral bands (*p*) (Hsu, 2007). This 'small *n* large *p* problem' has been termed the 'curse of dimensionality' which leads to the 'peaking phenomenon' or 'Hughes phenomenon' (Hsu, 2007). The 'Hughes phenomenon' introduces multi-collinearity in the input data matrix which makes the estimation of statistical parameters for the classifier performance inaccurate and unreliable (Kavzoglu and Mather, 2002; Hsu, 2007). Furthermore, computational requirements for processing large hyperspectral data sets might be prohibitive and time-consuming (Kavzoglu and Mather, 2002; Bajcsy and Groves, 2004). Therefore, techniques that reduce the 'curse of dimensionality' without sacrificing significant information are highly sought and feature selection or extraction task is often considered to be a practical and vital method in hyperspectral data processing and analysis (Shaw and Manolakis, 2002; Pal, 2005; Borges *et al.*, 2007).

Several hyperspectral feature or band selection techniques have been proposed to reduce the 'curse of dimensionality' and to identify the optimal bands required for discriminating and mapping wetland species (Daughtry and Walthall, 1998; Thenkabail *et al.*, 2002; Schmidt and Skidmore, 2003; Thenkabail *et al.*, 2004; Vaiphasa *et al.*, 2005; Vaiphasa *et al.*, 2007). These methods can be classified into the wrapper or filter approaches, based on whether or not they use classification algorithms as part of the evaluation process (Kavzoglu and Mather, 2002). The wrapper approach is a feature selection algorithm that searches for the best subset of bands using the classification algorithm as part of the evaluation process. On the other hand, the filter approach evaluates subsets of bands using the training data and without direct reference to the classification algorithm (Kohavi and John, 1997; Kavzoglu and Mather, 2002). The filter approach is computationally more efficient and has been more commonly used than the wrapper approach (Schmidt and Skidmore, 2003; Vaiphasa *et al.*, 2005; Ismail *et al.*, 2007). In the application of high dimensionality data such as hyperspectral data, it is recommended that the classification algorithm should be a part of the variable selection process (Granitto *et al.*, 2006). It is therefore desirable to have an algorithm that offers direct measuring of the importance of variables at the same time of the classification process of hyperspectral data (Ismail, 2009). This method is more efficient in several respects: 1. it uses the full available training data with no need for a validation set, 2. the method reaches a solution faster by avoiding retraining a predictor from scratch for every variable subset investigated (Guyon and Elisseeff, 2003).

Methods such as support vector machines, classification and regression trees, and neural networks have proved to be successful for the classification of hyperspectral data (Pal and Mather, 2004; Mutanga and Skidmore, 2004b; Questier *et al.*, 2005). However, the major shortcomings of support vector machines, classification and regression trees, and neural networks is that they lack any insight regarding the bands that best contribute to the derived classifier and are prone to overfitting and instability, the latter with particular reference to classification and regression trees (Archer and Kimes, 2008). Alternatively, RF (Breiman, 2001) is a bagging (bootstrap aggregation) operation where multiple classification trees are constructed based on a random subset of samples derived from the training data. The multiple classification trees then vote by plurality on the correct classification (Breiman, 2001; Lawrence *et al.*, 2006). Researchers have shown that this process decreases the correlation between the trees in the forest and yields an ensemble with low bias and low variance (Díaz-Uriarte and de Andrés, 2006; Archer and Kimes, 2008). Therefore, RF has many advantages over conventional classification tree-based approaches (Breiman, 2001). The stopping rules and pruning of trees is not necessary, and the approach has been shown to be robust to overfitting (Lawrence *et al.*, 2006). Overall, RF is relatively easy to implement when compared to the other ensemble classification methods and requires the user to specify only the (i) number of trees to be grown (*ntree*) and (ii) number of variables to split the nodes of individual trees (*mtry*) (Díaz-Uriarte and de Andrés, 2006). More importantly, studies have shown that RF can be successfully used for feature selection as well as for classification purposes (Svetnik *et al.*, 2003; Hamza and Larocque, 2005; Díaz-Uriarte and de Andrés, 2006; Granitto *et al.*, 2006; Han *et al.*, 2007; Archer and Kimes, 2008). However, only a few remote sensing studies have applied RF for feature selection and classification of hyperspectral data (Lawrence *et al.*, 2006; Chan and Paelinckx, 2008; Ismail, 2009).

Therefore, this study intends to investigate whether RF and canopy reflectance spectra resampled to HYMAP spectral resolution can discriminate amongst papyrus and its co-existing species in The Greater St Lucia Wetland Park. More specifically, the objectives of the study were to 1. Examine the utility of the RF wrapper based approach for selecting the optimal number of hyperspectral wavebands in a multiclass application, 2. Examine if the RF algorithm can accurately classify papyrus and its co-existent species in complex environments, where the vegetation classes have similar spectral characteristics and are affected by the underlying soils and hydrological regime, and 3. Examine the robustness of the RF algorithm in an application where the number of samples are less than the number of variables (*p*) (i.e., $n < p$).

4.2 Materials and methods

4.2.1 Spectral data acquisition and processing

Random points were generated on a land cover map that was derived from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) imagery. The sample points were subsequently uploaded into a GPS and used to navigate to the field sites i.e. Futululu Park, and the Mfabeni and Mkuzi swamps. Purposive sampling was done when the random point was not accessible, or to increase the variation of reflectance measurements. Once the sample site was located, a 3 m by 3 m vegetation plot was created to cover a homogenous area of the papyrus swamp or its co-existing species, and then the canopy spectral reflectance was measured.

All the spectral measurements were collected in December 2009 between 10:00 am and 02:00 pm under sunny and cloudless conditions using the Analytical Spectral Devices (ASD) FieldSpec® 3 spectrometer. The spectrometer measures wavelengths ranging from 350 nm to 2500 nm with a sampling interval of 1.4 nm for the 350 nm to1000 nm spectral region, and a 2.0 nm sampling interval for the 1000 nm to 2500 nm spectral region. The ASD has a spectral resolution of between 3 nm and 10 nm (ASD Analytical Spectral Devices Inc., 2005). A white reference spectralon calibration panel was used every 5 to 10 measurements to offset any change in the atmospheric condition and irradiance of the sun. Accompanying the field spectral measurements, metadata such as the sites' description (coordinates, altitude, and land cover class) and general weather conditions were also recorded (Milton *et al.*, 2009). Approximately 20 to 25 field spectrometer measurements were randomly taken at nadir from 1 m using a 5^o field of view (Table 4.1). This resulted in a ground field of view of about 18 cm in diameter, which was large enough to cover a cluster of papyrus and its co-existing species and reduce the background effects caused by soil and water (Mutanga *et al.*, 2004). These spectral measurements were then averaged to obtain the final spectral measurement for each vegetation plot.

Table 4.1: The number of sample plots and the total number of spectral measurements collected for papyrus and its associated species

Species name	Type code	Number of plots	Number of measurements
Cyperus papyrus	CP	55	1240
Phragmites australis	PA	53	1166
Echinochloa pyramidalis	EP	56	1288
Thelypteris interrupta	TI	51	1130

The spectral measurements from each of the wetland species (n = 4) were resampled to HYMAP spectra using ENVI 4.3 image processing software (Figure 4.1). The method used a Gaussian model with a full width at half maximum (FWMAP) equal to the band spacing provided (Mutanga, 2005). HYMAP is an airborne hyperspectral imaging spectrometer, comprising 126 wavelengths, operating over the spectral range between 436.5 nm – 2485 nm, with average spectral resolutions of 15 nm (437 nm -1313 nm), 13 nm (1409 nm – 1800 nm), and 17 nm (1953 nm – 2485 nm) (Cho *et al.*, 2007) . The spectral reflectance was resampled because the current operational airborne and spaceborne sensors such as HYMAP lack the fine spectral resolution of the ASD spectral reflectance (Mutanga, 2005). Additionally, in view of the current availability of airborne sensors in South Africa, it is of interest if the specific spectral bands of these sensors can discriminate between papyrus swamp and its co-existing species. If the results are positive, the mapping and monitoring of wetland plant species could be operational on airborne hyperspectral platforms.

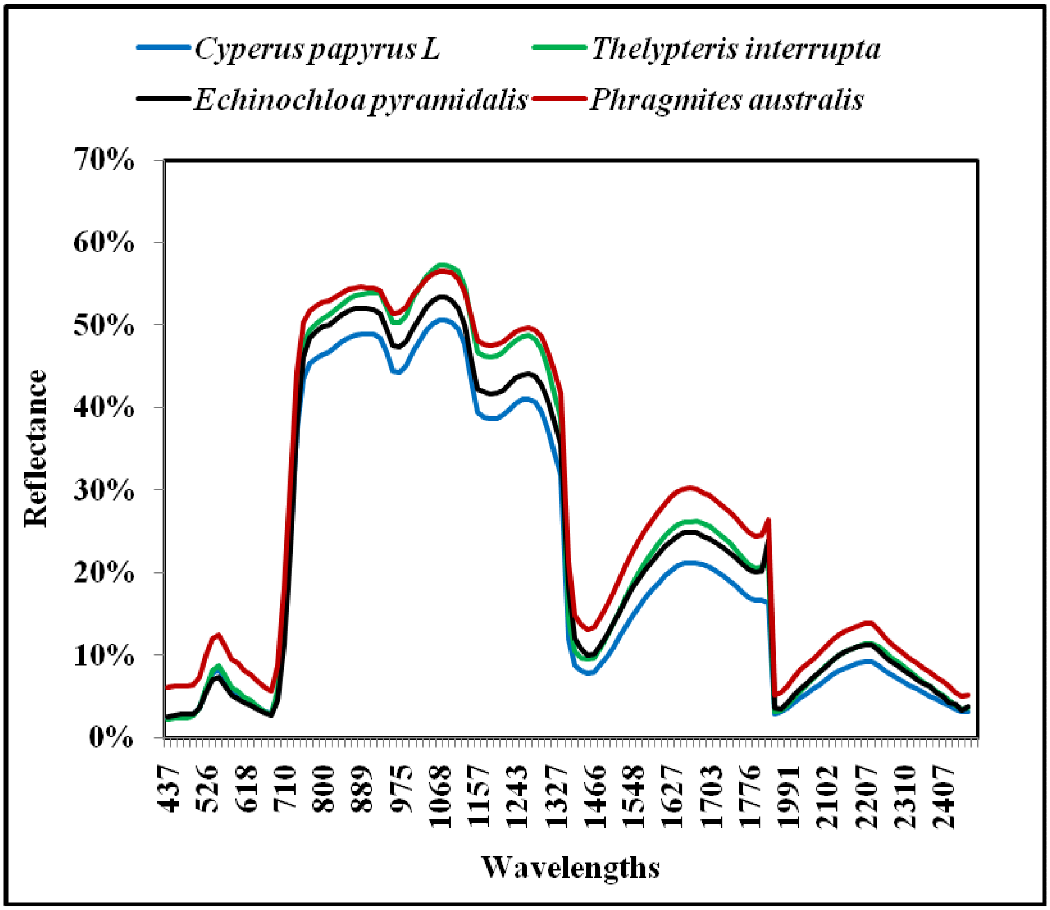

Figure 4.1. Mean canopy reflectance of resampled HYMAP data for *Cyperus papyrus L.* and its co-existing species: (Echinochloa *pyramidalis*, *Phragmites australis*, and *Thelypteris interrupta*).

4.2.2 Data analysis

4.2.2.1 Variables importance using the random forest algorithm

Random forest calculates three-variables importance measures, namely, the number of times each variable is selected, the Gini importance, and the permutation accuracy importance measure (Strobl *et al.*, 2007). The permutation of the variables, however, is considered to be the most advanced measure because of its ability to evaluate the variable importance by the mean decrease in accuracy using the internal out-of-bag (OOB) estimates while the forests are constructed (Breiman, 2001; Lawrence *et al.*, 2006; Strobl *et al.*, 2007).

In this study, we adopted the out-of-bag method to calculate the importance of a specific predictor variable (in our case wavelengths) in discriminating papyrus swamp and its co-existent species (Cutler *et al.*, 2007; Archer and Kimes, 2008; Chan and Paelinckx, 2008).The importance of each variable (n =126) used in this study was calculated based on how much worse the classification accuracy (mean decrease in accuracy) would be if that variable (wavelength) was permuted randomly (Prasad *et al.*, 2006). The importance of each variable is estimated in the following steps (i) the reflectance values of each wavelength is randomly permuted for the OOB samples, and then this modified OOB data are passed down each tree to get new predictions, (ii) the difference between the misclassification rate for the modified and original OOB data over all the trees that are grown in the forest are then averaged, (iii) this average is a measure of the importance of the variables and it is used as a ranking index which can be used to identify the wavelengths with relatively large importance in the classification process (Cutler *et al.*, 2007; Archer and Kimes, 2008; Chan and Paelinckx, 2008).

The optimization of the two parameters (*ntree* and *mtry*) of the RF algorithm is necessary to guarantee high accuracy of classification. In this regard, the big number of trees (*ntree*) is recommended to ensure that every input feature gets predicted several times (Kim *et al.*, 2006). We therefore optimized the *ntree* by using different values based on out-of-bag estimates of error (Liaw and Wiener, 2002). We also optimized the *mtry* number by trying all possible values (the default number is the square root of the number of variables). The RF library developed in R statistical software (R Development Core Team, 2007) was used to implement the RF algorithm.

4.2.2.2 Forward variable selection using the random forest algorithm

Although RF provides a measure of variables importance, it does not automatically choose the optimal number of variables that yield the best classification accuracy. The question therefore remains: What is the optimal number of wavelengths that can yield the smallest misclassification error rate? In this regard, we implemented a simple forward variable selection (FVS) method to identify the optimal subset of wavelengths with the lowest misclassification error. The FVS method uses the ranking of wavelengths as determined by the RF algorithm. The FVS method iteratively builds multiple random forests using the ranked wavelengths, and for each iteration five wavelengths were added to the model, and the error was calculated using OOB estimates error. Initially, the top five ranked wavelengths are selected and for the next iteration, the top 10 ranked wavelengths are selected. This process was repeated for the maximum amount of variables used in this study (n = 126). To validate the results from the OOB estimate error, we carried out a .632+ bootstrap method (n =100) (Díaz-Uriarte and de Andrés, 2006; Ismail, 2009). The .632+ bootstrap method uses a weighted average of the resubstitution error (the error when the RF classifier is applied to the training data) and the error on samples not used to train the predictor (the 'leave-one-out' error) (Díaz-Uriarte and de Andrés, 2006). The optimal subsets of wavelengths that yielded the smallest error rate as determined by the OOB method and .632+ bootstrap method were then used for classifying papyrus and its co- existent species.

4.2.2.3 Classification and accuracy assessment

It has been reported that, with the RF algorithm, it is not necessary to have cross validation or a separate accuracy assessment data set, because the OOB error provides an unbiased estimate of error (Lawrence *et al.*, 2006; Prinzie and Van den Poel, 2008). Therefore we used OOB to estimate the misclassification error. The confusion matrix was subsequently constructed to compare the true class with the class assigned by the classifier and to calculate the overall accuracy as well as the user and producer accuracy. Furthermore, a discrete multivariate technique called kappa analysis that uses the *k* (KHAT) statistic was also calculated to determine if one error matrix is significantly different from another (Cohen, 1960). This statistic serves as an indicator of the extent to which the percentage of correct values of an error matrix are due to the actual agreement in the error matrix and the chance agreement that is indicated by the row and column totals (Congalton and Green, 1999). If the KHAT coefficients are one or close to one then there is perfect agreement.

We also used the .632+ bootstrap method (n = 100) to estimate the misclassification error rate of the RF algorithm (Díaz-Uriarte and de Andrés, 2006; Granitto *et al.*, 2006). The .632+ bootstrap method was also applied to compare the error rate of RF with classification tree algorithms as an alternative method using the same data set. We used the 'errorest library' (Peters *et al.*, 2002) from the R statistical software (R Development Core Team, 2007) to calculate the .632+ bootstrap error.

4.3 Results

4.3.1 Effects of random forest input parameters on misclassification error

Before examining variable selection, it was essential to evaluate the effect of the user defined parameters (*mtry* and *ntree*) on the misclassification error. Figure 4.2 shows that the default setting of *mtry* (n = 11) proved to be the best choice in terms of the OOB error rate (11. 5 %). When examining the *ntree* parameter, results showed that the OOB error rates were relatively stable after 6000 trees (Figure 4.3), and we therefore used the default *mtry* and 6000 trees for *ntree* for all the further analyses.

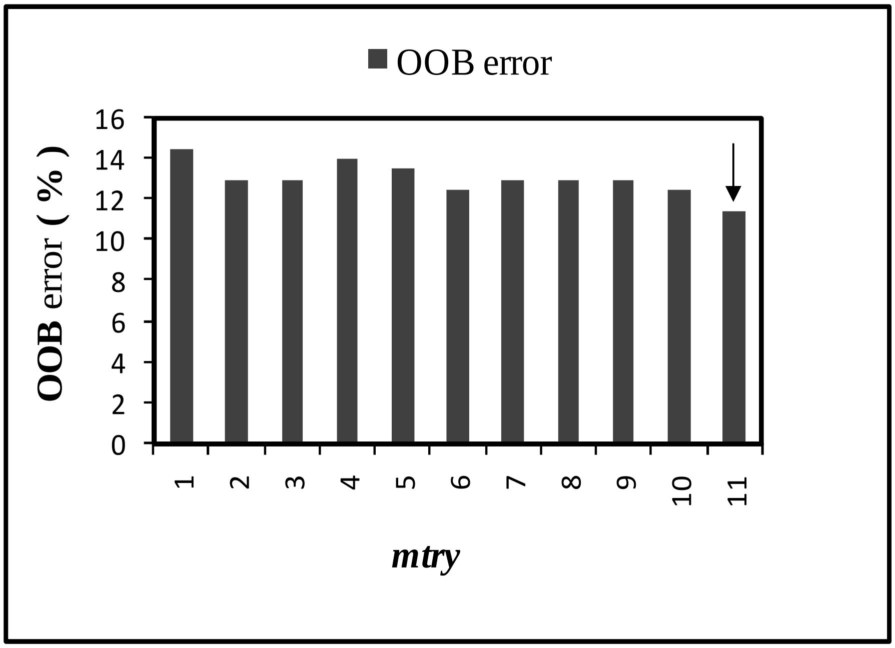

Figure 4.2. The effect of the number of variables tried at each split (*mtry*) on the performance of RF using the OOB estimate of error.

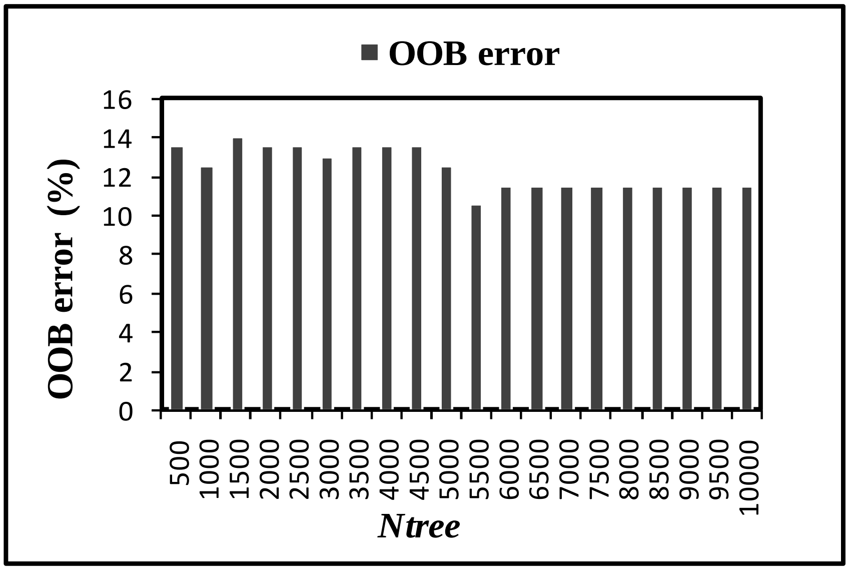

Figure 4.3. The effect of the number of trees (*ntree*) parameter on the performance of random forest using the OOB estimate of error (%).

4.3.2 Variables selection using the OOB method

All the resampled HYMAP wavelengths (n = 126) were used as input variables into the RF algorithm (default *mtry* value of 11 and 6000 trees (*ntree*). The RF algorithm yielded an OOB error rate of 11.5 % for the entire model. The mean decrease in accuracy as calculated by the OOB sample was then used to rank the wavelengths. Figure 4.4 shows the importance of all wavelengths as calculated by the RF algorithm. Results showed that the wavelengths with the highest mean decrease in accuracy are located predominately in the short wave infrared region (i.e.1409 nm and 1424 nm) and the visible region (i.e.710 nm and 437 nm). Additional important wavelengths are located between 437 nm and 710 nm.

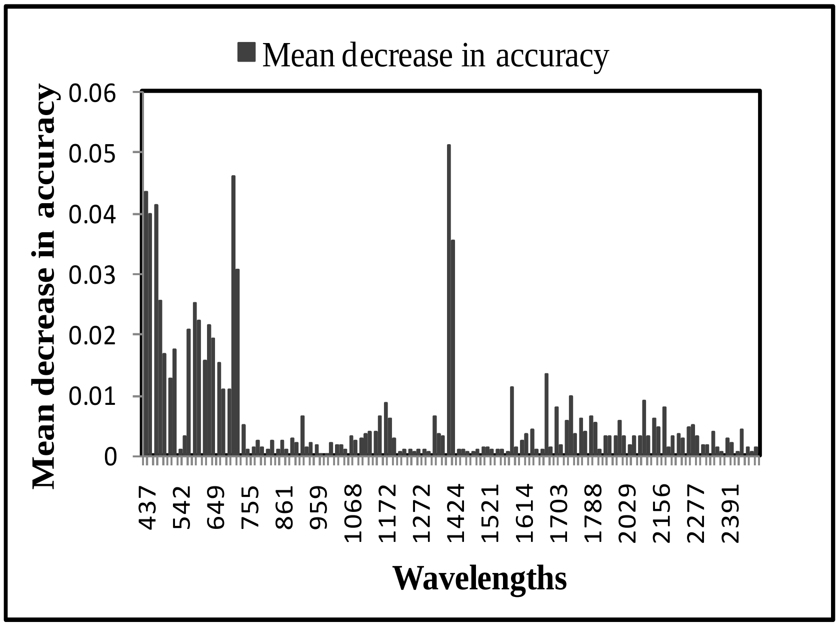

Figure 4.4. Variables importance as determined by the RF algorithm. The important wavelengths are those with the highest mean decrease in accuracy.

4.3.3 Forward variables selection method (FVS)

Figure 4.5 shows that the lowest misclassification rate as determined by both the .632+ bootstrap method (8.95 %) and the OOB method (9.5 %) is obtained when using 10 wavelengths located at 1409 nm, 710 nm, 437 nm, 464 nm, 452 nm,1424 nm,725 nm,480 nm, 587 nm, and 603 nm (the ranking is based on the importance measures). Using all wavelengths (n = 126) yielded a .632+ bootstrap error of 9.19 % and an OOB error estimate of 11.5 %. It is interesting to note that the OOB and the .632+ bootstrap error rates follow a similar trend (Figure 4.5). The top 10 bands were then used as input variables into the final RF model to classify papyrus swamp and its co-existent species.

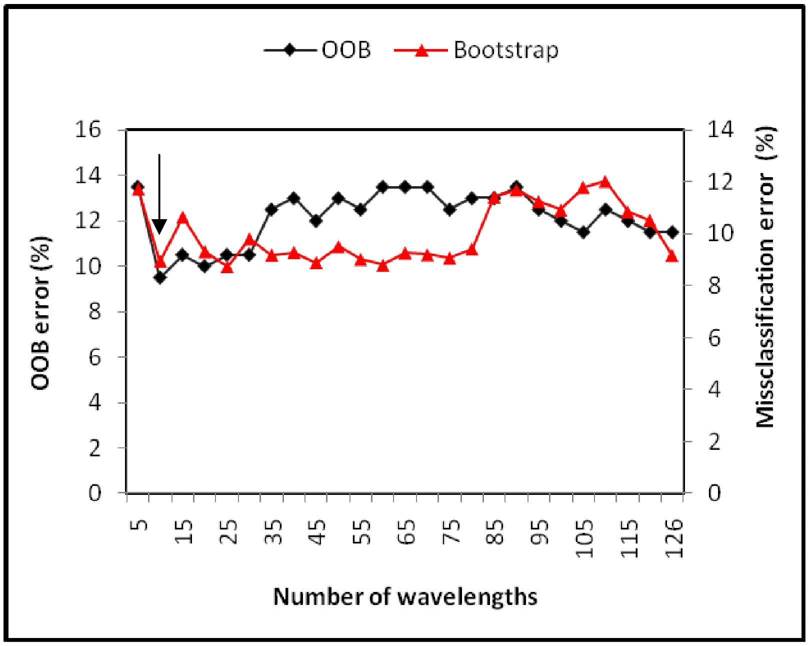

Figure 4.5. The forward variable selection method for identifying the optimal subset of wavelengths based on the OOB and .632+ bootstrap error estimates. The best subset of wavelengths with the lowest error rate is shown by the black arrow.

4.3.4 Classification and accuracy assessment

the results as shown in Table 4.2 indicate that the overall OOB error rate for all the classes was 9.5% using the ten wavelengths selected by the FVS method compared to the 11.5% obtained when all the wavelengths (n = 126) were used. For discriminating individual species, the confusion matrix shows that the *Phragmites australis* class has the lowest error rate (96 %), while the *Echinochloa pyramidalis* class has the highest error rate (86 %). Following the calculation of the overall OOB estimate of error, we subsequently used the confusion matrix shown in Table 4.2 to examine the error rate between papyrus and its co-occurring species. We examined the classification of each species (i.e. *Cyperus papyrus*, *Echinochloa pyramidalis*, *Phragmites australis* and *Thelypteris interrupta*) with every other species (Table 4.3).

Table 4.2: The confusion matrix showing the overall classification accuracy for *Cyperus papyrus L.* (CP), *Echinochloa pyramidalis* (EP), *Phragmites australis* (PA) and *Thelypteris interrupta* (TI).

	CP	EP	PA	TI	Class accuracy (%)
CP	45	1	1	3	90
EP	2	43	2	3	86
PA	1	1	48	0	96
TI	3	2	0	45	90
					Overall classification accuracy = 90.5%

Table 4.3 shows that the classification between the class pair of *Phragmites australis* and *Thelypteris interrupta* has a classification accuracy of 100%. The selected wavelengths (n = 10) also yielded a high classification accuracy between *Cyperus papyrus* and *Thelypteris interrupta* (97.89%) and *Echinochloa pyramidalis* (96.7%), and between *Cyperus papyrus* and *Phragmites australis* (93.75 %). *Thelypteris interrupta* appears to be unique amongst the other species based on the highest classification accuracy (94.62 % to 100 %) obtained. The overall classification accuracy obtained for all classes was 90.5 %. Table 4.3 also presents an overall KHAT value of 0.87 which indicates that there is strong agreement between the observations and the model predictions.

Table 4.3: The confusion matrix showing the classification error obtained for discrimination amongst all possible species combinations (n = 6). *Cyperus papyrus* (CP), *Echinochloa pyramidalis* (EP), *Phragmites australis* (PA), and *Thelypteris interrupta* (TI) . The confusion matrix includes the accuracy between classes (ACC), the KHAT statistic, producer accuracy (PA), and user accuracy (UA)

Classes	ACC %	KHAT	PA %		UA %		Row totals	Column totals
			Presence	Absence	Presence	Absence		
CP *vs* EP	96.70	0.93	95.74	97.73	97.83	95.56	91	91
CP *vs* TI	97.89	0.96	97.83	97.96	97.83	97.96	95	95
CP *vs* PA	93.75	0.88	93.75	93.75	93.75	93.75	96	96
EP *vs* PA	96.81	0.94	97.73	96.00	95.56	97.97	94	94
EP *vs* TI	94.62	0.89	95.56	93.75	93.48	95.74	93	93
PA *vs* TI	100.00	1.00	100.00	100.00	100.00	100.00	93	93
All classes	90.50	0.87	88.24	91.49	90.00	86.00	200	200

We used the .632+ bootstrap method to compare the performance of the RF algorithm against the widely used classification trees (CT) algorithm (Harb *et al.*, 2009; Ismail and Mutanga, 2009). The results of performance assessments are shown in Figure 4.6 for both machine learning methods using different subsets of wavelengths. It is clear, over a range of different subsets of wavelengths used, the overall misclassification error rates obtained by the RF algorithm are much lower than the misclassification error rates obtained by the CT algorithm. It is interesting that the use of the top 10 wavelengths (1409 nm, 710 nm, 437 nm, 464 nm, 452 nm, 1424 nm, 725 nm, 480 nm, 587 nm, and 603 nm) yielded the lowest misclassification error rate for both the CT algorithm (15.05 %) and RF (9.5 %) and that the highest misclassification error was obtained using the top 5 wavelengths for the RF algorithm (11.74 %) and CT (18.56 %).

We also used the confusion matrix to compare the KHAT values and overall accuracy between the two machine learning methods. Table 4.4 shows that the RF model produces better overall accuracy and KHAT value compared to classification trees algorithm for all HYMAP wavelengths and the top 10 wavelengths. The overall accuracy and KHAT values for the RF algorithm were 90.5 % and 87 %, and for the CT algorithm were 84.5 % and 80 % respectively when the top 10 wavelengths were used. The RF algorithm also yielded better classification accuracy (88.44%) than the CT algorithm (80.47%) when the full data set (126 wavelengths) was used.

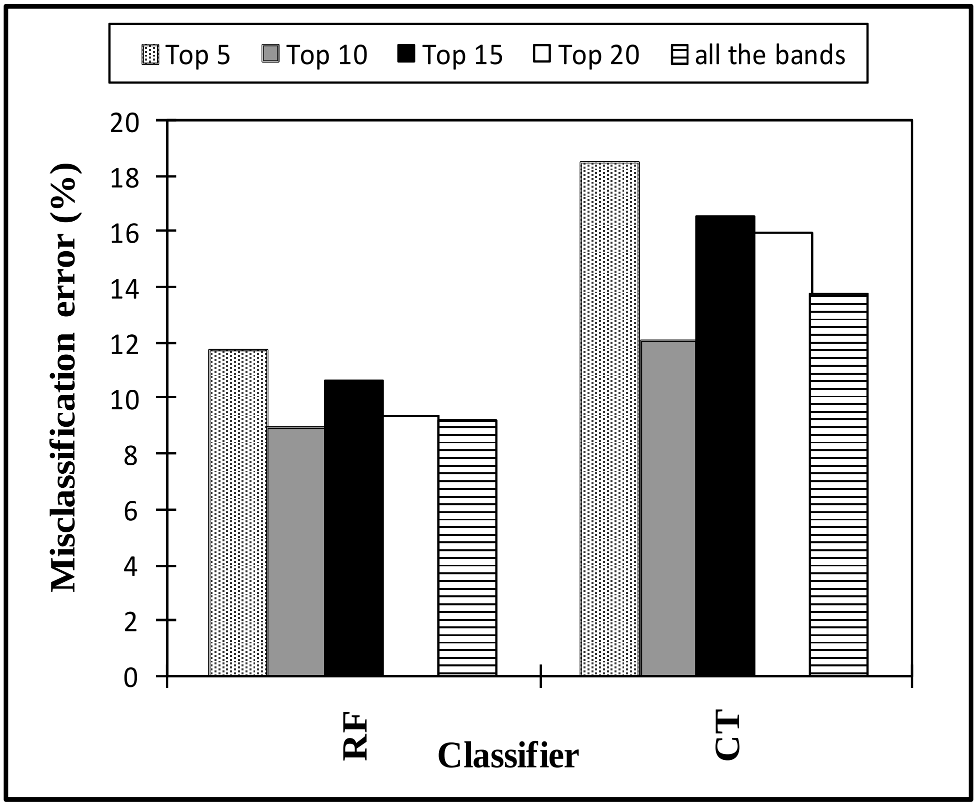

Figure 4.6. Comparison between the performance of the random the forest algorithm (RF) and the classification tree algorithm (CT) using different subsets of wavelengths selected by RF. The misclassification error rate was estimated using the .632+ bootstrap (n= 100).

Table 4.4: The misclassification error for both machine learning models (RF and CT) using the .632+ bootstrap method for error estimates and the accuracy assessments using the top 10 wavelengths selected by the RF algorithm and a full data set (126 wavelengths).

Algorithm	Top 10 wavelengths			Full data set		
	Misclassification error %	Overall Accuracy	KHAT %	Misclassification error	Overall accuracy	KHAT %
RF	8.95	90.5	87	9.19	88.44	85
CT	12.05	84.5	80	13.75	80.47	78

4.4 Discussions

This study tested the utility of field spectrometry data resampled to HYMAP resolution and the RF algorithm for variables selection and classification of *Cyperus papyrus* and its co-existent species (*Phragmites australis*, *Echinochloa pyramidalis*, and *Thelypteris interrupta*) located in the St Lucia Wetlands Park, South Africa.

Overall, the results obtained in this study show the benefit of using the RF algorithm for identifying key wavelengths as well as for producing excellent classification results. Additionally, results show that when optimizing the RF algorithm the default setting of *mtry* (in our case *mtry* = 11) is sufficient. These results are identical to those of Díaz-Uriarte and de Andrés (2006) and Liaw and Wiener (2002) who indicated the insensitivity of RF to the choice of *mtry* and found that the highest accuracy can be achieved by using a large number of trees. With this range of capabilities, RF offers powerful alternatives to traditional parametric and semi-parametric statistical methods for the analysis of hyperspectral data.

However, the limitation of the RF algorithm was that it does not automatically choose the optimal number of variables that could yield the lowest error rate. The FVS method used in this study provided the optimal numbers of important variables (n = 10) that offer the lowest misclassification error rate. Results show that the full HYMAP data set (n=126) produced an overall accuracy of 88.44 % and a KHAT value of 0.85 compared to when the selected wavelengths (n = 10) were used producing an overall accuracy of 90.5 % and a KHAT value of

0.87. Using the selected wavelengths produced a 2 % increase in classification accuracies. These results obtained are comparable to those of Lawrence *et al.* (2006) who found that using a full data set of Probe-1 (128 bands) with the RF algorithm in classifying two invasive species produced lower overall accuracy than when the data set was reduced. These results emphasize the assertion that, in the model-based analysis, the increase of hyperspectral bands could lead to a decrease in the classification accuracy because the noise in the redundant data propagates through the classification model (Benediktsson *et al.*, 1995; Bajcsy and Groves, 2004).Therefore, the use of large and redundant numbers of hyperspectral bands (in our case n = 126) has resulted in lower classification accuracy (overall accuracy 88.44 %) and (OOB estimate error 11.5 %) than processing a subset of relevant bands (in our case n = 10) without redundancy (overall accuracy 90.5 %) and OOB estimate error (9.5 %). Overall, the result shows the excellent performance of the FVS method in dimensionality reduction without sacrificing significant spectral information.

Previous studies that have classified wetland vegetation using remotely sensed data have shown the relative importance of the visible infrared (VIS) and short wave infrared (SWIR) in discriminating wetland species (Daughtry and Walthall, 1998; Schmidt and Skidmore, 2003; Thenkabail *et al.*, 2004; Vaiphasa *et al.*, 2005). Similarly, the ten wavelengths selected in this study (1409 nm, 710 nm, 437 nm, 464 nm, 452 nm, 1424 nm, 725 nm, 480 nm, 587 nm, and 603 nm) emphasized the potential usefulness of the visible region and SWIR even at a coarser HYMAP spectral resolution in discriminating the wetland species . However, the results produced higher classification accuracies when compared to research carried out by Pengra *et al.* (2007) who achieved an overall accuracy of 81.4 % for mapping *Phragmites australis* using EO-1 Hyperion hyperspectral sensor. In this study we obtained a classification accuracy of 96% for *Phragmites australis.* However, it should be noted that the results of this study are based on resampled HYMAP data. Noise in the blue part of the spectrum and atmospheric absorption, especially around the selected bands (1409 nm and 1424 nm) might present some problems when upscaling the results to an airborne platform. We believe that the techniques used in this study should receive considerable additional testing with other airborne or spaceborne data. Nevertheless, the results from this study demonstrate the possibility of hyperspectral data to map papyrus and its co-existent species in swamp wetlands.

We compared the utility of the RF algorithm against the widely used tree-based ensemble classifier (classification trees algorithm) using the .632+ bootstrap estimate error. Our evaluation criteria included accuracy assessment using all HYMAP wavelengths and wavelengths selected by RF. In these experiments, we found that the RF algorithm obtained higher overall classification accuracy than the CT algorithm for all the wavelengths and selected wavelengths. The results obtained in this study are consistent with those of Hamza and Larocque (2005) and Pal (2005) who showed that the RF algorithm achieves the best classification accuracy compared to other ensemble methods that use tree classifiers as the base model. It is also interesting to note that wavelengths selection by RF (n = 10) produced lower misclassification error (12.05 %) for the CT algorithms than other different subsets of HYMAP wavelengths (Figure 4.6). This result emphasizes the robustness and reliability of RF as a variables selection method and for producing the best classification accuracy.

4.5 Conclusions

This paper aimed at discriminating *Cyperus papyrus, Phragmites australis, Echinochloa pyramidalis,* and *Thelypteris interrupta* located in the Greater St Lucia Wetlands Park, South Africa, using RF and field spectrometry data resampled to HYMAP sensor.

Our results have shown that:

1. The proposed method for variables selection was able to provide small sets of non-redundant wavelengths while preserving highest classification accuracy.

The study demonstrated the possibility to scale up the method to airborne sensors such as HYMAP for discriminating swamp wetland species with an overall accuracy of 90.5 %. We believe that the techniques used in this study should receive considerable additional testing with other airborne or spaceborne data.

2. Based on relatively high accuracy, low cost (availability of R statistics package is free of charge), simplicity, and few parameters to be set, RF algorithms can be considered as a new approach for the analysis of hyperspectral data.

Overall, the results from this study have revealed that RF algorithm is a robust and accurate method for the combined purpose of variables selection and for the classification of hyperspectral data in an application where (i) the number of samples is limited (n < p), and where (ii) vegetation species have similar spectral characteristics affected by underlying wet soil and hydrology regime. However, further studies are necessary for testing the stability and reliability of the internal assessment of accuracy (OOB) in the RF algorithm using an independent accuracy assessment data set. Given the problem of soil and water background affecting the spectral reflectance of papyrus and other species, it would be useful if the use of vegetation indices in discriminating these species is further investigated in future studies.

Acknowledgements

We thank the Greater St Lucia Wetlands Park authorities for allowing us access to the study areas. The contributions of Khalid Manssour in the fieldwork are gratefully acknowledged. We extend our gratitude to Rick Lawrence for his thoughtful comments on facilitating the task. Our thanks go to Susan Davies for her help in proofreading. We thank the University of KwaZulu-Natal for providing the optimal conditions for the success of the study and assistance with the ASD. We thank Elfashir University for its contribution in funding this study.

CHAPTER FIVE

Improving spectral discrimination of papyrus (*Cyperus papyrus L.*) and its co-existing species using narrow band vegetation indices

This chapter is based on:

Adam, E., Mutanga, O., and Ismail, R., (in review). Improving the spectral discriminating of papyrus (*Cyperus papyrus L.)* and its co-existent species at canopy level with hyperspectral indices and random forest algorithm. *International Journal of Applied Earth Observation and Geoinformation.*

Abstract

Recent advances in hyperspectral remote sensing provide opportunities to discriminate and map wetland plant species. The shortcoming of individual spectral bands for discriminating wetland plant species is their limited spectral information which might be inadequate for characterizing the spectral reflectance of wetland vegetation which is highly correlated and combined with the reflectance spectra of the underlying wet soil and hydrologic regime. The objective of this study is to evaluate the potential of hyperspectral vegetation indices for improving the spectral discrimination of papyrus and three co-existing species in the Greater St Lucia Wetland Park-South Africa. *In situ* canopy reflectance measurements ranging from 350 nm to 2500 nm were taken from papyrus and the three co-existent species using an analytical spectral devices spectrometer. We calculated the normalized difference vegetation index (NDVI) and a simple ratio (SR) involving all possible two-band combinations of the 20 most important bands as determined by the RF algorithm. In addition, we evaluated a number of hyperspectral indices (n = 48) that were previously demonstrated to estimate plant parameters such as biomass, leaf area index (LAI), chlorophyll *a* and *b*, and nitrogen concentration. An analysis of variance and a simple forward variable selection technique were used to select optimal vegetation indices that showed the highest potential to discriminate between the wetland species. Three of the optimal vegetation indices were previously published in the literature (Plant Senescence Reflectance Index, Blue/Green Index 1, and Pigment Index 4) while the other two optimal indices were obtained from the modified NDVIs involving a combination of a narrow band in the red portion (655 nm) with two wavelengths in the red-edge position (697 nm and 705 nm). Finally, the RF algorithm was used to classify the species using the optimal indices. An overall accuracy of 96% was obtained using the out-of-bag data with individual class accuracies ranging from 93.7 % to 100 %. Our results clearly indicate that: 1. hyperspectral indices might offer new possibilities of discriminating plant species, and 2. the out-of-bag data, as an internal estimate of accuracy in the RF algorithm, provide a reliable and stable accuracy assessment so it might be unnecessary to have independent accuracy assessment data when using the RF algorithm.

Keywords: Field spectrometer measurements. Papyrus. Random forest. Vegetation indices. Variable selection. Wetland vegetation.

5.1 Introduction

Mapping and monitoring wetland species such as papyrus (*Cyperus papyrus L.*) are key requirements for a better understanding of the functions and dynamics of wetland ecosystems and are also critical for effective and sustainable management of wetlands (Schmidt and Skidmore, 2003; Zomer *et al.*, 2009). Quantitative, accurate, and repeatable techniques for discriminating wetland vegetation at species level in large areas are therefore of self-evident importance (Mutanga *et al.*, 2003; Belluco *et al.*, 2006; Lawrence *et al.*, 2006). Traditional survey methods such as hand mapping and Global Positioning Systems (GPS) receiver mapping have proven to be highly accurate for small management areas (Cooksey and Sheley, 1997). However, these methods require intensive fieldwork ,including taxonomical information, collateral and ancillary data analysis, and the visual estimation of percentage cover for each species, which might be economically, technically, and logistically inadequate for wetland environments because of their high diversity and poor accessibility (Xie *et al.*, 2008; Zomer *et al.*, 2009).

Hyperspectral remote sensing provides opportunities to discriminate and map wetland vegetation at species level due to the availability of narrow spectral channels of less than 10 nm (Schmidt and Skidmore, 2003; Rosso *et al.*, 2005; Vaiphasa *et al.*, 2005; Belluco *et al.*, 2006; Vaiphasa *et al.*, 2007; Zomer *et al.*, 2009). These narrow bandwidth data permit an in-depth detection of detailed vegetation species which could otherwise be masked when using the broad wavebands of the Landsat TM or SPOT sensors (May *et al.*, 1997; Harvey and Hill, 2001; McCarthy *et al.*, 2005). However, while the high spectral resolution of hyperspectral data facilitates accurate detection and identification, the high dimensionality of the data causes substantial problems in analysing and processing complexity (Demir and Ertürk, 2008). Additionally, redundancy in the hyperspectral data exists due to strong correlation between adjacent spectral bands (Jiang *et al.*, 2004). The calculation of narrow band vegetation indices offers a suitable method to overcome the problems of high dimensionality and redundancy in the hyperspectral data (Andrew and Ustin, 2006).

Additionally, the use of hyperspectral technology for discriminating wetland plant species is challenging because the reflectance spectra of wetland vegetation canopies, especially in the visible and near-infrared region, are highly correlated due to their similar biochemical and biophysical properties (Price, 1992; Kokaly *et al.*, 2003; Adam and Mutanga, 2009).

Furthermore, when these properties are combined with reflectance spectra of the underlying soil and hydrologic regime (Yuan and Zhang, 2006), there is a decrease in the spectral reflectance, especially in the near- to mid-infrared regions where water absorption is relatively stronger (Fyfe, 2003; Silva *et al.*, 2008). Therefore, information in a single spectral band might be inadequate to characterize the vegetation properties or to identify the factors affecting their spectral reflectance (Zhao *et al.*, 2007).

In the past three decades, several spectral vegetation indices (VIs) have been developed to provide more sensitive measurements of plant biophysical parameters such as biomass, LAI, water content, and chlorophyll (Green *et al.,* 1997; Sims and Gamon, 2002; Mutanga and Skidmore, 2004a; Stimson *et al.*, 2005; Zhao *et al.*, 2007; Darvishzadeh *et al.*, 2008; Xue and Yang, 2009) to reduce external noise interferences such as those related to soil, atmosphere condition, and sun view angles (Mutanga and Skidmore, 2004a) and to enhance the variability of spectral reflectance of vegetation (Qi *et al.*, 1995; Haboudane *et al.*, 2002; Cho *et al.*, 2008). These VIs were developed mathematically based either on narrow band spectral data or broad band sensor such as Landsat TM and SPOT. Studies have shown that these VIs provide more highly correlated relationships with vegetation properties than individual bands (Tanriverdi, 2006).

The normalized difference vegetation index (NDVI) (Tucker, 1979), and simple ratio (SR) (Maxwell, 1976) are the most commonly used broad band indices in correlating remote sensing observations with the characteristics of vegetation (Zhao *et al.*, 2007; Cho *et al.*, 2008). NDVI calculation is based on the contrasting intense chlorophyll absorption in the red (670 to 680 nm) against the high signal in the near-infrared wavelength (750 nm to 850 nm) due to light scattering by leaves (Mutanga and Skidmore, 2004a; Cho *et al.,* 2008). NDVI calculated from broad band sensors has been found useful in classifying wetlands at coarse levels (Johnston and Barson, 1993) and estimating biomass (Tan *et al.*, 2003) and LAI of wetland vegetation (Green *et al.*, 1997). However, attempts to discriminate vegetation species have not been possible because they produce similar NDVI values (Nagendra, 2001). Furthermore, the limitation of standard vegetation indices constructed with red and near-infrared spectral measurements, particularly NDVI, is that they yield poor estimates after a certain biomass density or LAI due to the saturation problem (Mutanga and Skidmore, 2004a).

Hyperspectral indices, on the other hand, have been shown to be significantly correlated with biochemical and physiological properties of vegetation canopies and leaves. The concentrations of these biochemical and physiological properties depend on factors such as phenology, degree of canopy development, and type of environment stress (Cho *et al.*, 2008). Therefore, the difference in physiological concentrations of vegetation can be another source of variation in plant spectral signatures (Best *et al.*, 1981; Silva *et al.*, 2008). Because hyperspectral indices developed in the visible and near-infrared region respond to these differences in physiological status and environmental factors (Silva *et al.*, 2008), therefore, it might offer the possibility to map plant species or communities depending on their differences in canopy structures and biochemical compositions (Nagendra, 2001; Cho *et al.*, 2008).

Studies involving variation in plant spectral signatures based on different phenological stages and physiological and biochemical concentrations have been conducted in a single species or plant community and have not been carried out between different plant species and communities (Best *et al.*, 1981; Mutanga *et al.*, 2003). Most of these studies investigated the use of the novel spectral indices derived from leaf scale measurements and have rarely the indices examined for species discrimination (Cho *et al.*, 2008). The canopy spectra indices have shown better plant species discrimination as compared to leaf spectra indices using visible and near-infrared wavelengths (400 nm to 900 nm) (Cho *et al.*, 2008).

Methods such as support vector machines, classification and regression trees, and neural networks have proved to be successful for the classification of hyperspectral data (Pal and Mather, 2004; Mutanga and Skidmore, 2004b; Questier *et al.*, 2005). However, the major shortcomings of support vector machines, classification and regression trees, and neural networks is that they lack any insight regarding the bands that best contribute to the derived classifier and are prone to overfitting and instability, the latter with particular reference to classification and regression trees (Archer and Kimes, 2008). Alternatively, the RF algorithm (Breiman, 2001) is a bagging (bootstrap aggregation) operation where multiple classification trees are constructed based on a random subset of samples derived from the training data. The multiple classification trees then vote by plurality on the correct classification (Breiman, 2001; Lawrence *et al.*, 2006). Researchers have shown that this process decreases the correlation between the trees in the forest and yields ensemble with low bias and low variance (Díaz-Uriarte and de Andrés, 2006; Archer and Kimes, 2008). Therefore, the RF algorithm has many advantages over conventional classification tree-based approaches (Breiman, 2001). The stopping rules and pruning of trees is not necessary, and the approach has been shown to be robust to overfitting (Lawrence *et al.*, 2006). Overall, the RF algorithm is relatively easy to implement when compared to the other ensemble classification methods and requires the user to specify only the (i) number of trees to be grown (*ntree*) and (ii) number of variables to split the nodes of individual trees (*mtry*) (Díaz-Uriarte and de Andrés, 2006). More importantly, studies have shown that RF can be successfully used for feature selection as well as for classification purposes (Svetnik *et al.*, 2003; Díaz-Uriarte and de Andrés, 2006; Granitto *et al.*, 2006; Han *et al.*, 2007; Archer and Kimes, 2008). However, only a few remote sensing studies have applied the algorithm for feature selection and classification of hyperspectral data (Lawrence *et al.*, 2006; Chan and Paelinckx, 2008; Adam *et al.*, 2009; Adam *et al.*, In press).

Therefore, the objective of this study was to explore the performance of various hyperspectral vegetation indices derived from canopy scale measurements in discriminating papyrus and three other co-existent species. We selected and computed band ratios which have been widely and successfully used in vegetation studies (Mutanga and Skidmore, 2004a). We examined narrow band NDVI and SR involving all possible two-band combinations of the top 20 bands measured with the RF algorithm. Some of the existing hyperspectral indices (n = 48) that were previously demonstrated to estimate plant parameters such as biomass, LAI, chlorophyll a and b, and nitrogen concentration were also considered. Further, we tested the reliability of the internal accuracy assessment of the RF algorithm using an independent accuracy assessment data set.

5.2 Materials and methods

5.2.1 Canopy spectral measurements

Random points were generated on a land cover map that was derived from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) imagery. The sample points were subsequently uploaded into a GPS and used to navigate to the field sites i.e. Futululu Park, and the Mfabeni and Mkuzi swamps. Once the sample site was located, a 3 m by 3 m vegetation plot was created to cover a homogenous area of the papyrus swamp or its co-existing species, and canopy spectral reflectance was then measured.

All the spectral measurements were collected in December 2009 between 10:00 am and 02:00 pm under sunny and cloudless conditions using the Analytical Spectral Devices (ASD) FieldSpec® 3 spectrometer. The spectrometer measures wavelengths ranging from 350 nm to 2500 nm with a sampling interval of 1.4 nm for the 350 nm to1000 nm spectral region, and 2.0 nm sampling interval for the 1000 nm to 2500 nm spectral region. The ASD has a spectral resolution between 3 and 10 nm (ASD Analytical Spectral Devices Inc., 2005). A white reference spectralon calibration panel was used every 5 to 10 measurements to offset any change in the atmospheric condition and irradiance of the sun. Accompanying the field spectral measurements, metadata such as the sites' description (coordinates, altitude, and land cover class) and general weather conditions were also recorded (Milton *et al.*, 2009). Approximately 20 to 25 field spectrometer measurements were randomly taken at nadir from 1 m and using a 5^o field of view (Table 5.1). This resulted in a ground field of view of about 18 cm in diameter, which was large enough to cover a cluster of papyrus and its co-existing species and reduce the background effects caused by soil and water (Mutanga *et al.*, 2004). These spectral measurements were then averaged to obtain the final spectral measurement for each vegetation plot.

Table 5.1: The number of sample plots and the total number of spectral measurements collected for papyrus and its associated species

Species name	Type code	Number of plots	Number of measurements
Cyperus papyrus	CP	82	1476
Phragmites australis	PA	83	1328
Echinochloa pyramidalis	EP	86	1688
Thelypteris interrupta	TI	80	1130

We randomly divided the spectral data for each species into two equal data sets (Lawrence *et al.*, 2006), and the models were developed using one-half of the data (n = 40), while the withheld half of the data (n = 40) was used for an independent accuracy assessment.

5.2.2 Vegetation indices calculation

Two types of hyperspectral indices were tested in this study: 1. Hyperspectral indices that were previously demonstrated to estimate plant parameters such as biomass, LAI, chlorophyll a and b, and nitrogen concentration (Table 5.2), and 2. Narrow band vegetation indices computed according to the principle of the NDVI (Eq.1) and SR (Eq.2) from all possible two-band combination indices involving 20 bands selected by the RF algorithm. This resulted in 800 indices, 400 NDVIs and 400 SRs.

$$NDVI = \frac{R_{(i,n)} - R_{(j,n)}}{R_{(i,n)} + R_{(j,n)}} \qquad (1)$$

$$SR = \frac{R_{(i,n)}}{R_{(j,n)}} \qquad (2)$$

Where $R_{(i,n)}$ and $R_{(j,n)}$ are the reflectance of any two bands from the 20 bands selected by the RF algorithm for each species.

Table 5.2: Vegetation indices used in this study

No	Index name	Abbreviation	Formula*	References
1	Normalized different vegetation index	NDVI	(R830-R670)/ (R830+R670)	(Rouse *et al.*, 1974)
2	Carter index	CI	R760/R695	(Carter, 1994)
3	Gitelson and Merzylak Index	GMI	R750/R700	(Gitelson and Merzlyak, 1994)
4	Vogelman index	VOG	R740/R720	(Vogelmann *et al.*, 1993)
5	Photochemical reflectance index	PRI	(R531-R570)/ (R531+R570)	(Penuelas *et al.*, 1995)
6	Normalized Difference	ND	(R750-R705) /(R750+R705)	(Gitelson and Merzlyak, 1994)
7	Structure Insensitive Pigment Index	SIPI	(R800-R445) /(R800-R680)	(Penuelas *et al.*, 1995)
8	Pigment Specific SR (chlorophyll a)	PSSRa	R800/R680	(Blackburn, 1998)
9	Pigment Specific SR (chlorophyll b)	PSSRb	R800/R635	(Blackburn, 1998)
10	Simple Ratio 1	SR1	R695/R420	(Carter, 1994)
11	Simple Ratio 2	SR2	R695/R760	(Carter, 1994)
12	Plant Senescence Reflectance Index	PSRI	(R680-R500)/R750	(Merzlyak *et al.*, 1999)
13	Simple Ratio 3	SR3	R750/R710	(Gitelson and Merzlyak, 1994)
14	Modified Chlorophyll Absorption in Reflectance Index	MCARI1	[(R700-R670)-0.2(R700-R550)] (R700/R670)	(Daughtry *et al.*, 2000)
15	Transformed Chlorophyll Absorption in Reflectance Index	TCARI	3[(R700-R670)-0.2(R700-R550)(R700/R670)]	(Haboudane *et al.*, 2002)

Table 5.2: Vegetation indices used in this study (cont.)

No	Index name	Abbreviation	Formula*	References
16	Optimized Soil-Adjusted Vegetation Index	OSAVI	(1+0.16)(R800-R670)/(R800+R670+0.16)	(Rondeaux *et al.*, 1996)
17	Modified Chlorophyll Absorption in Reflectance Index	MCARI2	1.2[2.5(R800-R670)-1.3(R800/R550)]	(Haboudane *et al.*, 2002)
18	Anthocyanin Reflectance Index 1	ARI 1	(1/R550)-(1/R700)	(Gitelson *et al.*, 2001)
19	Anthocyanin Reflectance Index 2	ARI 2	R800[(1/R550)-(1/R700)]	(Gitelson *et al.*, 2001)
20	Blue/Green Index	BGI 1	(R400)/(R550)	(Zarco-Tejada *et al.*, 2005)
21	Blue/Green Index	BGI 2	(R450)/(R550)	(Zarco-Tejada *et al.*, 2005)
22	Carotenoid Reflectance Index 1	CRI 1	(1/R510)-(1/R550)	(Gitelson *et al.*, 2002)
23	Carotenoid Reflectance Index 2	CRI 2	(1/R510)-(1/R700)	(Gitelson *et al.*, 2002)
24	Modified Red Edge Normalized Difference Vegetation Index	MNDVI 705	(R750-R705)/(R750 +R705-2R445)	(Sims and Gamon, 2002)
25	Modified Red-Edge Simple Ratio Index	MSR 705	(R750-R445)/(R705-R445)	(Sims and Gamon, 2002)
26	Moisture Stress Index	MSI	R1599/R819	(Hunt and Rock, 1989)
27	Water Band index	WBI	R900/R970	(Penuelas *et al.*, 1997)
28	Normalized Difference Water Index	NDWI	(R857-R1241)/(R857 +R1241)	(Gao, 1996)
29	Ratio Analysis of Reflectance Spectra	RARSa	R675/R700	(Chappelle *et al.*, 1992)
30	Ratio Analysis of Reflectance Spectra	RARSb	R675/(R650R700)	(Chappelle *et al.*, 1992)
31	Ratio Analysis of Reflectance Spectra	RARSc	R760/R500	(Chappelle *et al.*, 1992)
32	Pigment Specific Simple Ratio	PSSRa	R800/R680	(Blackburn, 1998)
33	Normalized Difference Vegetation Index	NDVI	(R813-R613)/ (R813+R613)	(Ma *et al.*, 2001)
34	Green Normalized Difference Vegetation Index	GNDVI	(R875-R560)/ (R875+R560)	(Penuelas *et al.*, 1995)
35	Normalized Pigment Chlorophyll Ratio Index	NPCI	(R680-R430)/ (R680+R430)	(Gitelson and Merzlyak, 1996)
36	Structurally Independent Xanthophylls Index	SIXI	(R430-R800)/ (R680+R800)	(Penuelas *et al.*, 1995)
37	Soil Adjusted Vegetation Index	SAVI	1.5(R780-R670) /(R780+R680+0.5)	(Huete, 1988)
38	Photochemical Reflectance Index	PRI	(R531-R570)/ (R531+R570)	(Rahman *et al.*, 2001)

Table 5.2: Vegetation indices used in this study (cont.)

No	Index name	Abbreviation	Formula*	References
39	Red/Green Ratio	RG	(R600-R699)/ (R500-R599)	(Fuentes *et al.*, 2001)
40	Simple Ratio Pigment Index	SRPI	R430/R680	(Zarco-Tejada, 1998)
41	Normalized Phaeohytinization index	NPQI	(R415-R435)/ (R415+R435)	(Zarco-Tejada, 1998)
42	Structure Intensive Pigment Index	SIPI	(R800-R445)/ (R800-R680)	(Zarco-Tejada, 1998)
43	Pigment Index 1	PI 1	R695/R420	(Zarco-Tejada, 1998)
44	Pigment Index 2	PI 2	R695/R760	(Zarco-Tejada, 1998)
45	Pigment Index 3	PI 3	R440/R690	(Lichtenthaler *et al.*, 1996b)
46	Pigment Index 4	PI 4	R440/R740	(Lichtenthaler *et al.*, 1996b)
47	Normalized Difference Nitrogen Index	NDNI	log (R1680/R1510)/ Log (1/R1680 R1510)	(Serrano *et al.*, 2002)
48	Normalized Difference Lignin Index	NDLI	log (R1680/R1754)/ Log (1/R1680 R1754)	(Serrano *et al.*, 2002)

*R= reflectance measurements

5.2.3 Statistical analysis

5.2.3.1 The random forest algorithm

The random forest algorithm is a modified bagging (bootstrap aggregation) classifier where multiple classification trees are developed, and the final classification is determined by a majority vote. Each tree in the forest is trained on a bootstrapped sample (i.e. 2/3 of the original observations) (Breiman, 2001), and at each node of individual trees, the RF algorithm searches only across a random subset of the variables (i.e. spectral indices) to determine the split. The trees are then grown to maximum size without any pruning (Breiman, 2001; Lawrence *et al.*, 2006). Additionally, RF has an intrinsic means to estimate variable importance and to assess accuracy by using the out-of-bag data. Researchers have commented that a separate test data set may not be required for accuracy assessments (Lawrence *et al.*, 2006).The out-of-bag error estimates (1/3 of the original data) are created from the data that are not in the bootstrap sample used for each tree's development (Breiman, 2001). To guarantee high accuracy of classification, studies have recommended that the two parameters of RF have to be optimized; these parameters are the number of trees (*ntree*) grown in a forest and the number of variables (*mtry*) used in each tree split (Breiman, 2001; Liaw and Wiener, 2002; Díaz-Uriarte and de Andrés, 2006; Archer and Kimes, 2008). The default value of *mtry* is the square root of the number of variables (Liaw and Wiener, 2002). However, a large *ntree* value and default value of *mtry* are recommended (Gislason *et al.*, 2006; Kim *et al.*, 2006; Adam *et al.*, 2009). We, therefore, developed the model using an *ntree* value of 10000, and the default number of *mtry*. The RF library (Liaw and Wiener, 2002), developed in R statistical software (R Development Core Team, 2007), was used to implement the RF algorithm.

5.2.3.2 Variables importance using the random forest algorithm

The RF algorithm calculates three variable importance measures, namely, the number of times each variable is selected, the Gini importance, and the permutation accuracy importance measure (Strobl *et al.*, 2007). The permutation of a variable, however, is considered to be the most advanced measure because of its ability to evaluate the variables importance by the mean decrease in accuracy using the internal out-of-bag (OOB) estimates while the forests are constructed (Breiman, 2001; Lawrence *et al.*, 2006; Strobl *et al.*, 2007).

In this study, we used mean decrease in accuracy using the internal OOB estimates (Cutler *et al.*, 2007; Archer and Kimes, 2008; Chan and Paelinckx, 2008). The importance of each variable (wavelength) used in this study was calculated based on how much worse the classification accuracy (mean decrease in accuracy) would be if the data of that variable were permuted randomly (Prasad *et al.*, 2006). To calculate the importance of wavelength in discriminating papyrus and its co-existent species, the reflectance values of each wavelength were randomly permuted for the OOB data, and then the modified OOB data were passed down the tree to get a new classification. The difference between the misclassification rate for the modified and original out-of-bag data over all the trees grown in the forest are then averaged. The average, which is therefore a measure of the importance of the variable is used as a ranking index (Cutler *et al.*, 2007; Archer and Kimes, 2008; Chan and Paelinckx, 2008), that can be used to identify the wavelengths with relatively large important scores for the calculation all possible two-band combination indices. The top 20 wavelengths that showed the highest importance based on mean decrease in accuracy were subsequently selected for calculating the all possible two-band combination indices (Guyon and Elisseeff, 2003).

5.2.3.3 Variables selection: Filter approach

To assess the potential of the various VIs used in this study for species discrimination a one-way analysis of variance (ANOVA) was used as a filter approach to test if the differences in the spectral indices of papyrus and the other three species were statistically significant. In this regard, the research hypothesis is that the spectral indices between each class pair of the species (CP,PA, EP, and TI) were significantly different , the null hypothesis Ho: μ1 = μ2 = μ3 = μ4 versus the alternate hypothesis Ha: μ1 ≠ μ2 ≠ μ3 ≠ μ4 where: μ1, μ2, μ3, and μ4 are the spectral indices values from *Cyperus papyrus L* (CP), *Phragmites australis* (PA), *Echinochloa pyramidalis* (EP), and *Thelypteris interrupta (*TI) respectively. We tested ANOVA with a 99% confidence level ($p < 0.01$). Furthermore , a Tukey's HSD post hoc test was carried out in order to determine if there was a difference in the mean between the various class pairs (i.e. CP *vs.* PA, CP vs. EP, CP vs. TI, PA vs. EP, PA vs. TI, and EP vs. TI). Histogram and matrix plots were then used to examine which indices could most frequently discriminate all the species. VIs with no statistical significance were then discarded, while the significant indices for all class pairs were retained for further analysis.

5.2.3.4 Optimal subset of vegetation indices

Using an ANOVA with Tukey's HSD post hoc test is limited because the method does not automatically select the optimal subset of VIs that have the strongest discriminatory power. In other words the method examines each VI individually as opposed to considering interaction between VI's. The question, therefore, remains: What is the optimal number of significant vegetation indices that can yield the smallest misclassification error rate? In this regard, we applied a forward variable selection using the RF algorithm to identify the optimal subset of VIs (Guyon and Elisseeff, 2003). The RF algorithm was used to compute and rank the importance of each significant VI in discriminating the species. The method involves iteratively fitting multiple random forests (on the training data) and at each iteration building a RF after sequentially adding the indices with the highest important values. Initially, the top ranked vegetation index is selected, and for the next iteration the top two ranked indices are added and so on. The error for each iteration is then calculated using the OOB samples. The procedure was repeated for the maximum number of significant vegetation indices used in this study. The optimal subset of indices which yielded the smallest out-of-bag error was then used for classifying papyrus and its co- existent species.

5.2.3.5 Classification accuracy

To evaluate the prediction performance of an algorithm, the use of a large independent test data set that has not been used in the training is recommended (Congalton and Green, 2008). However, when the data are limited some types of cross-validation techniques are usually carried out (Hawkins *et al.*, 2003). In the RF algorithm, the OOB estimate of error is considered to be such a type of cross-validation technique (Breiman, 2001). Specifically, at each bootstrap iteration a single tree is grown using a particular bootstrap sample. Since bootstrapping is sampling with the replacement from approximately two-thirds of the training data (in our case spectral indices), some of the variables will be left out of the sample and may not be used at all in any growing tree, while some others will be chosen more than once (Breiman, 2001; Svetnik *et al.*, 2003). The variables that have not been used in the tree growing constitute the OOB and are then used to estimate the prediction performance of the classifier (Breiman, 2001). In this study, we used the OOB method as internal estimate of error using the one-third portion of the data that was randomly excluded from the construction of each of the classification trees used. A confusion matrix was subsequently constructed to compare the true class with the class assigned by the classifier and to calculate the overall accuracy as well as the user and producer accuracy. Furthermore, a discrete multivariate technique called kappa analysis that uses the *k* (KHAT) statistic was also calculated to determine if one error matrix is significantly different from another (Cohen, 1960; Mutanga, 2005). This statistic serves as an indicator of the extent to which the percentage of correct values of an error matrix are due to the actual agreement in the error matrix and the chance agreement that is indicated by the row and column totals (Congalton and Green, 2008). If the kappa coefficients are one or close to one then there is perfect agreement between the observed and predicted class.

Lawrence *et al.,* (2006) recommended further testing for the reliability of OOB as an internal accuracy assessment of the RF classifier. Therefore, we used an independent test data set and the OOB samples to assess the classification accuracy. The OOB accuracy assessment of the training data was then compared to the accuracy of the predications obtained when using the test data set.

5.3 Results

5.3.1 Measuring the variables importance using the random forest algorithm

The RF algorithm was used to measure the importance of individual wavelengths (n = 1706). These wavelengths yielded an OOB error rate of 14.5 %. The mean decrease in accuracy as calculated by the OOB sample was then used to rank the wavelengths (Figure 5.1). Results clearly show that the top 20 wavelengths with the highest mean decrease in accuracy are located predominately in the red-edge portion (655 nm, 690 nm, 697 nm, 703 nm, 705 nm, 709 nm, 713 nm, 712 nm, 715 nm, 719 nm, 720 nm, and 721 nm), the near-infrared region (1337 nm, 1341 nm, 1347 nm, 1350 nm, and 1538 nm), and mid-infrared (2203 nm, 2198 nm, and 2199 nm). The top 20 wavelengths were then used to classify the species and yielded a lower OOB error rate of 8.5 %. In order to determine if VIs could yield a lower OOB error, we subsequently used these wavelengths to compute all the possible two-band NDVI and SR combinations.

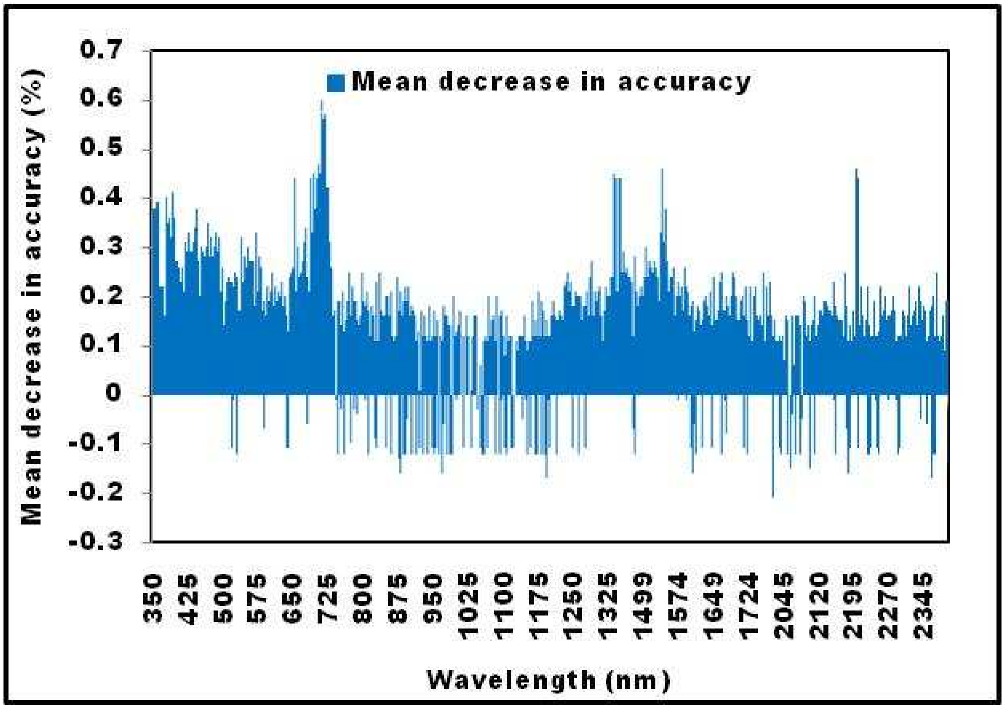

Figure 5.1. Variables importance as determined by the RF algorithm for 1706 wavelengths. The important wavelengths are those with the highest mean decrease in accuracy.

5.3.2 Variables selection using filter approach (ANOVA)

The top 20 wavelengths identified by the RF algorithm allowed for the computation of 400 narrow band NDVIs and 400 narrow band SRs. These narrow band indices (n = 800) as well as vegetation indices published in the literature (n = 48) were statistically analysed to test the hypothesis that the mean values of the vegetation indices used to discriminate between papyrus and the three other co-existent species were significantly different. Results of the one-way ANOVA indicate that there is a statistically significant difference among the species ($p < 0.01$). The results of the multiple comparisons between the class pairs (CP vs. PA, CP vs. EP, CP vs. TI, PA vs. EP, PA vs. TI, and EP vs. TI) using all possible two-band combinations are shown in (Figure 5.2). Figure 5.3 shows the results when using the published vegetation indices.

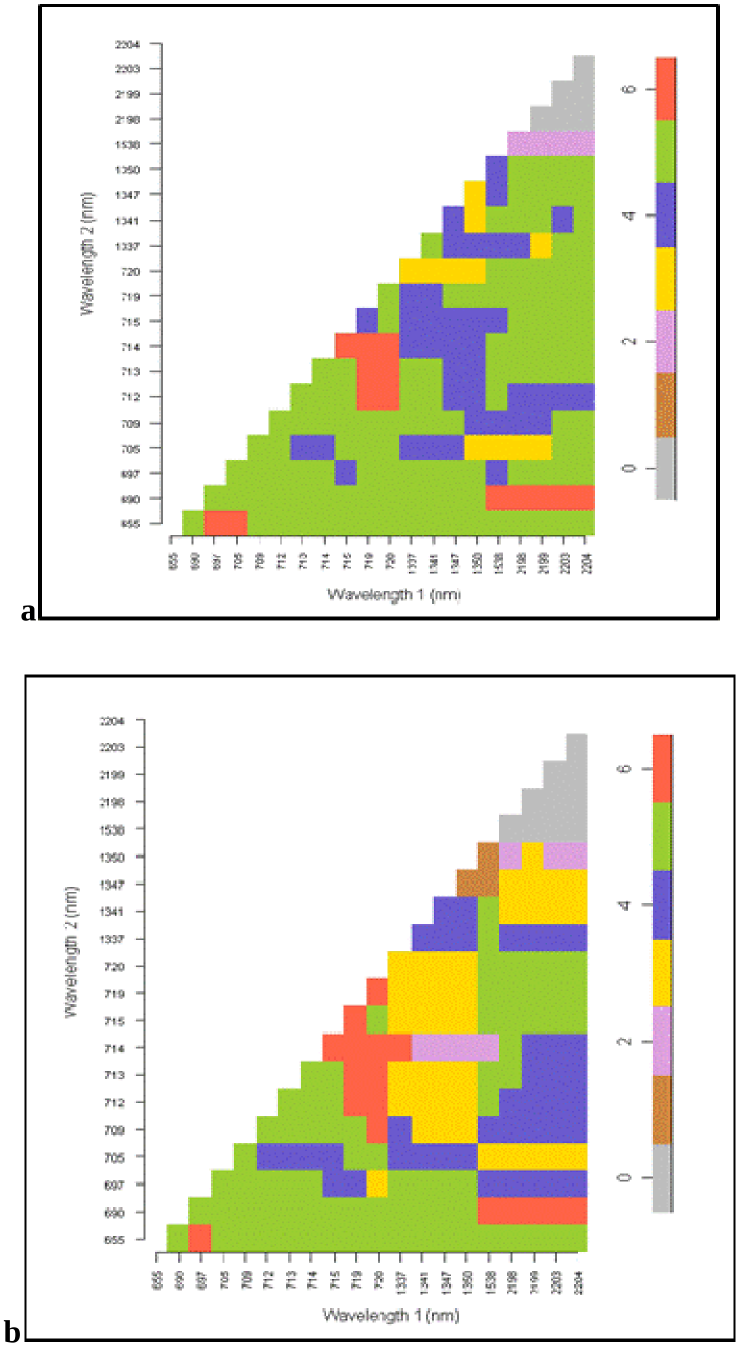

Figure 5.2. Matrix plots show the significant difference, marked by presence of a colour, of class pairs (n = 6) in each narrow band NDVI (a) and narrow band SR (b) that were calculated from all possible combinations involving the top 20 bands. The red colour indicates the vegetation indices that could discriminate between all class pairs (n = 6) of the species.

Figure 5.2 clearly shows that the majority of the VIs (NDVI and SR) that could discriminate between all class pairs (n = 6) were calculated on all possible two-band combinations located in red-edge portion (714 nm, 719 nm, 720 nm,712 nm, 713 nm, and 714 nm). Additionally, VIs that could discriminate between all class pairs were computed from a red-edge wavelength located at 690 nm combined with wavelengths located in the water absorption part of the spectrum (2203 nm, 2204 nm, 2198 nm, and 2199 nm) and a near-infrared wavelength located at (1538 nm). In total, 27 NDVI and 28 SR narrow band indices could discriminate between all class pairs (n = 6). With respect to the published vegetation indices, that included GMI, ND, SRI, PSRI, ARI1, ARI2, BG1, BG2, CRI2, RARSb, PI1, PI3, and PI4 (Figure 5.3) are most successful in discriminating between all class pairs.

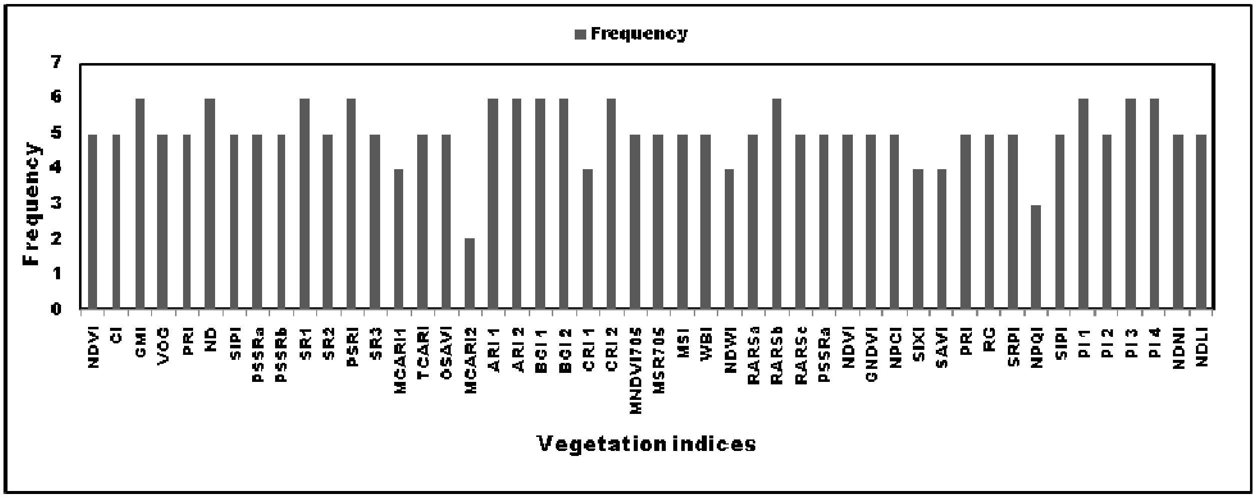

Figure 5.3. Frequency of statistically significant differences for all class pairs (CP vs. PA, CP vs. EP, CP vs. TI, PA vs. EP, PA vs. TI, and EP vs. TI). The maximum frequency number (6) indicates the vegetation indices that could discriminate between all class pairs (n = 6) of the species.

5.3.3 Forward variables selection

First, we used OOB to estimate the error rate of different combinations among the significant indices (NDVIs, SRs, and published VIs) to retain the indices that yielded the smallest error for forward variable selection. The OOB error rate is shown in Table 5.3.

Table 5.3: The OOB error rate for significant vegetation indices considered in this study. The random forest was built using default of *mtry* and an *ntree* value of 10000

Significant Vegetation indices	Number of significant indices	OOB error (%)
published VIs	13	15.5
Narrow band NDVIs	27	14
Narrow band SRs	28	16.5
Narrow band NDVIs and SRs	55	14
Narrow band NDVIs, SRs, and published VIs	68	12

As seen in Table 5.3, the combination involving the narrow band NDVIs, narrow band SRs, and VIs published in the literature yielded the lowest OOB error (12 %). Therefore, forward variable selection was carried out on this combination of VIs (n = 68) to select the optimal subset of VIs with strong discriminatory power for further classification. The optimal subset with the smallest error is shown in Figure 5.4.

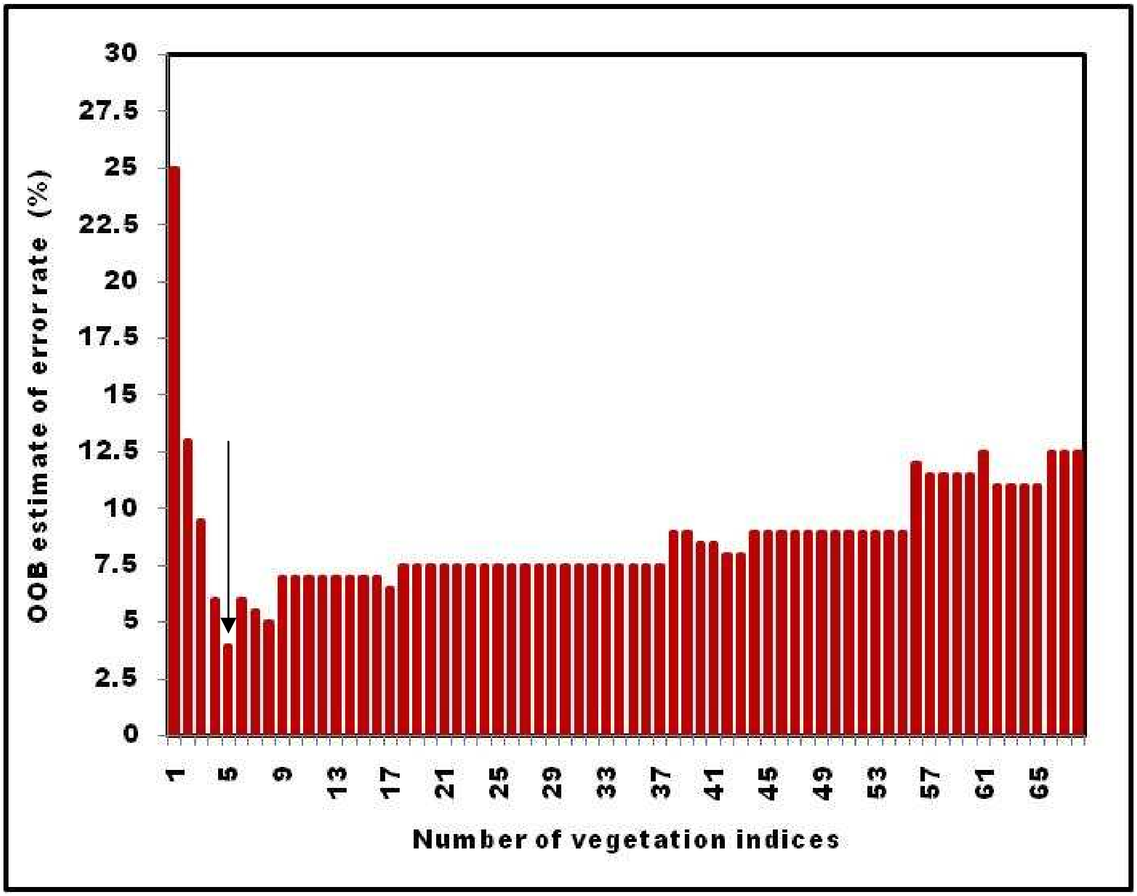

Figure 5.4. The forward variable selection method for identifying the optimal subset of vegetation indices using the OOB estimate of error rate. The best subset of vegetation indices with the lowest error rate is shown by the black arrow.

The results of the forward selection process indicate that a subset consisting of only five vegetation indices yielded the lowest OOB error (4 %). Three of these vegetation indices are from the VIs published in the literature (PSRI, BGI1, and PI4), and the other two vegetation indices are derived from the two-band combination (narrow band NDVIs) involving a wavelength located in the red portion (655 nm) combined with two wavelengths located in the red-edge position (697 nm and 705 nm). These vegetation indices (n = 5) were then retained for further classification. Noticeable in Figure 5.4 was that the OOB error increases when the numbers of vegetation indices increase.

5.3.4 Classification assessment

The optimal subset of vegetation indices (n = 5) was used as input variables in the RF classifier to discriminate papyrus and its co-existing species. An overall accuracy of 96% (*k* = 0.91) was obtained for all class pairs (CP vs. PA, CP vs. EP, CP vs. TI, PA vs. EP, PA vs. TI, and EP vs. TI) as determined by the OOB estimate of error rate. Additionally, the producer's accuracy for the class pairs ranged from 95 % for *Cyperus papyrus L.* and *Echinochloa pyramidalis* to 100 % for *Cyperus papyrus L.* and *Thelypteris interrupta* (Table 5.4). Utilizing all the significant indices (n = 68) produced an overall accuracy of 88% (*k* = 0.84) as estimated by the OOB estimate of error rate (Table 5.4). It is also interesting to note from Table 5.4 that all class pairs which involve *Cyperus papyrus L* (CP) had the highest class accuracies (93.7 % to 99 %). Overall results indicate that the best discrimination of *Cyperus papyrus L.* from its co-existing species is possible with the selected vegetation indices (n = 5). The performance of the out-of-bag estimate of accuracy was compared with that of an independent dataset using the optimal subset of vegetation indices (n = 5) and full data set (n = 68).Table 5.4 shows the results obtained from the two accuracy assessment methods.

Table 5.4: Accuracies assessment for OOB estimates and independent test data set based on the top five vegetation indices and the full data set (n = 68). The assessment includes the kappa statistic, overall accuracy (ACC), producer accuracy (PA), and user accuracy (UA).

	Top five vegetation indices								Full data set (68 vegetation indices)							
	Out-of-bag accuracy assessment				Independent accuracy assessment				Out-of-bag accuracy assessment				Independent accuracy assessment			
Classes	ACC %	Kappa	PA %	UA %	ACC %	Kappa	PA %	UA %	ACC %	Kappa	PA %	UA %	ACC %	Kappa	PA %	UA %
CP vs EP	93.7	0.87	95.7	91.7	94.4	0.89	92.6	96.2	92.2	0.84	95.4	89.1	98	0.96	96.2	100
CP vs TI	99.0	0.98	100	97.8	93.3	0.86	100	86.2	98.9	0.98	100	97.6	89.8	0.80	92.6	86
CP vs PA	99.0	0.98	100	97.8	100	1.00	100	100	98.3	0.83	89.1	93.2	94.3	0.89	92.6	96
EP vs PA	100	1.00	100	100	100	1.00	100	100	100	1.00	100	100	100	1.00	100	100
EP vs TI	95.9	0.92	97.8	93.8	96.6	0.93	100	92.9	91.3	0.83	95.4	87.2	92.6	0.85	100	86
PA vs TI	100	1.00	100	100	100	1.00	100	100	100	1.00	100	100	94.6	0.89	90.3	100
All classes	96.0	0.91	97.0	89	94.5	0.91	93.6	84.3	88.0	0.84	85.0	82.0	85.8	0.81	83.0	83.0

5.4 Discussion

This study aimed at discriminating papyrus vegetation (*Cyperus papyrus L.*) and three other co-existing species (*Phragmites australis, Echinochloa pyramidalis,* and *Thelypteris interrupta*) that dominate the swamp wetland of the GSWP. The motivation for the study was to investigate the possibility of using the RF algorithm and hyperspectral indices to improve discrimination among vegetation species in a swamp wetland that exhibits a complex ecosystem and hydrology regime.

5.4.1 Variables importance using the random forest algorithm

Hyperspectral data are very rich in information. However, the large number of highly correlated hyperspectral wavelengths poses many challenges such as the computational requirement, redundancy removal, and model accuracy assessment. Variables ranking is an effective technique to select a fixed number of top ranked variables of hyperspectral data for better classification (Pal, 2006). The results of this study confirm that the RF algorithm is an efficient method of ranking wavelengths (Figure 5.1) and allows focusing on a small subset of wavelengths (n= 20) for calculating the vegetation indices (NDVI, SR) from all possible two-band combinations (Figure 5.2). These top 20 wavelengths (655 nm, 690 nm, 697 nm, 703 nm, 705 nm, 709 nm, 713 nm, 712 nm, 715 nm, 719 nm, 720 nm, 721 nm, 1337 nm, 1341 nm, 1347 nm, 1350 nm, 1538 nm, 2203 nm, 2198 nm, and 2199 nm) are within ± 10 nm from known wavelengths that have been used in some other studies to discriminate wetland species. These are 1409 nm, 725 nm, and 710 nm (Adam *et al.*, 2009), 720 nm (Daughtry and Walthall, 1998; Thenkabail *et al.*, 2002; Vaiphasa *et al.*, 2005), and 705 nm (Thenkabail *et al.*, 2004). Moreover, the remarkable accuracy (96%) achieved in this study proved that this method is an effective procedure for calculating vegetation indices involving possible combinations between hyperspectral bands, and it also helps in the reduction of data dimensionality and therefore might be valuable in terms of data processing and analysis rather than handling all the data (350 nm to 2500 nm) which is difficult to compute and to select the relevant information

5.4.2 Significant difference in vegetation indices between the species

A one-way ANOVA with Tukey's HSD post hoc test was used to determine 1. whether there were statistically significant differences in VIs values among the four vegetation species, and 2. whether it could be used as a baseline filter approach for limiting the total number of vegetation indices (n = 848). When comparing the two plots of NDVI and SR in Figure 5.2, it is interesting to note that the VI's that could discriminate all class pairs (n= 6) were obtained by combining narrow bands located in shorter wavelengths of the red-edge portion of electromagnetic spectrum (712 nm, 713 nm, 714 nm, 719 nm, and 720 nm), and a wavelength located in the red-edge (690 nm), and the wavelengths of mid-infrared region (1538 nm, 2198 nm, 2199 nm, and 2203 nm). It is also interesting to note that most of the significant differences in vegetation indices published in the literature for full class pairs were obtained by combining narrow bands from the shorter wavelengths of the red-edge portion (700 nm to 760 nm). This included the VIs such as GMI, ND, SR2, RARSb, and PI2. VIs calculated from these shorter wavelengths of the red-edge portion are sensitive to variations in chlorophyll content and green biomass (Lichtenthaler *et al.*, 1996a; Mutanga and Skidmore, 2004a).

The differences in green wavelength peak (550 nm) have been used to successfully discriminate vegetation species characterized by differences in chlorophyll content (Peña-Barragán *et al.*, 2006). This has been confirmed in this study by the results obtained by the ANOVA that show that there is a high significant difference between all class pairs when using vegetation indices such as ARI1, ARI2, BGI1, and BGI2. These VIs were calculated using 550 nm (green peak) with combinations of wavelengths located at 400 nm, 450 nm, 700 nm, and 800 nm. It is therefore assumed that the hyperspectral difference between the four species (*Cyperus papyrus L*, *Phragmites australis, Echinochloa pyramidalis,* and *Thelypteris interrupta*) may be attributed to significant variation in the relative amount of chlorophyll content and green biomass. This is supported by the assertion that wetland plant species appear to vary greatly in chlorophyll content and biomass (Anderson, 1995). This variation is considered to be one of the variables affecting the spectral properties of papyrus and its co-existent species (Adam and Mutanga, 2009).

5.4.3 Optimal vegetation indices

Given that there are statistically significant differences (p< 0.001) in VIs values among the four vegetation species, what remains to be discovered is the optimal subset of significant vegetation indices that can yield smallest misclassification error. Results from this study confirm that the combination of forward variable selection and the RF algorithm is a useful approach to identify the most important or information-rich vegetation indices, thereby allowing the significant vegetation indices (n = 68) to be reduced in size (n =5). Our results as presented in Figure 5.4 show that five selected vegetation indices can discriminate among all the class pairs (CP vs. PA, CP vs. EP, CP vs. TI, PA vs. EP, PA vs. TI, and EP vs. TI) with a 4 % OOB error rate in comparison to 12 % OOB error rate obtained when utilizing all the VIs (n = 68). The results obtained in this study are comparable to other studies that revealed that the subsets of variables selected by the RF algorithm have produced higher overall accuracy than utilizing the full data set (Lawrence *et al.*, 2006; Adam *et al.*, 2009; Ismail, 2009). These results emphasize the assertion that, in the model-based analysis, the increase of hyperspectral variables could lead to a decrease in the classification accuracy because the noise in the redundant data propagates through the classification model (Benediktsson *et al.*, 1995; Bajcsy and Groves, 2004).

Overall, the result shows the excellent performance of the forward variable selection method applied in dimensionality reduction without sacrificing significant spectral information. Hence, classifying papyrus and its co-existing species can be made on the basis of these optimal vegetation indices (n = 5) to provide the highest classification accuracy.

5.4.4 Classification assessment

The estimated overall accuracy from the OOB estimate of error rate for optimal vegetation indices was 96 % (kappa = 0.91). These results are particularly remarkable when compared to a study by Adam *et al.* (2009) who used the RF algorithm and spectrometry data resampled to HYMAP resolution to classify the same species. Their study yielded an overall accuracy of 90.5% using 14 bands. This clearly shows that the overall accuracy has been improved in this study with 5.5 % using only a small subset of VIs (n = 5). The class accuracy was also improved, for example; our results produced higher classification accuracies when compared to research carried out by Pengra et *al.* (2007) who achieved an overall accuracy of 81.4 % for mapping *Phragmites australis* using EO-1 Hyperion hyperspectral sensor. In our study we obtained a classification accuracy of 99% to 100 % for the class pairs involving *Phragmites australis.*

Our results in this study confirmed the power of the RF algorithm in providing highest classification accuracy (more than 90 %) of hyperspectral data (Lawrence *et al.*, 2006; Pal, 2006; Adam *et al.*, 2009). It also shows the ability of vegetation indices in enhancing the possible difference in reflectance between the vegetation species (Qi *et al.*, 1995; Haboudane *et al.*, 2002; Peña-Barragán *et al.*, 2006; Cho *et al.*, 2008).

Our evaluations of the reliability of the out-of-bag estimates of accuracy as an internal method for accuracy assessment in RF have shown that this estimate is reliable and stable, especially with a high number of classification trees. This can be clearly seen in Table 5.4 that shows that the independent accuracy assessment is nearly identical to the OOB accuracy assessment. In this aspect, our result is consistent with that obtained by Lawrence *et al.* (2006) who found that the accuracy assessment using OOB is nearly identical to an independent accuracy assessment. Our results strengthen the assertion that with the RF algorithm it is not necessary to have a separate accuracy assessment if the reference data are protected against any type of bias (Lawrence *et al.*, 2006; Prinzie and Van den Poel, 2008). We believe that this study is protected against bias with a simple random sampling method applied for the reference data collection (Lawrence *et al.*, 2006; Congalton and Green, 2008).

5.5 Conclusions

This study aimed at improving discriminating *Cyperus papyrus L., Phragmites australis, Echinochloa pyramidalis,* and *Thelypteris interrupta* located in the Greater St Lucia Wetland Park, South Africa, using the RF algorithm and hyperspectral indices derived from field spectrometry data.

Our results have shown that:

1- The proposed method for ranking variables importance for possible two-band combinations and optimal subset of vegetation indices for species discriminating was efficient in providing small sets of data while preserving highest classification accuracy.

2- The optimal subset of vegetation indices that yielded the highest classification accuracy is sensitive to the variation in chlorophyll content and green biomass. Since these biochemical and biophysical variables were not measured in this study, it therefore remains to be explained why the selected vegetation indices showed a relatively better ability to discriminate between the species.

3- Based on relatively high overall accuracy (96 %), the use of hyperspectral indices can be considered as a new approach for discriminating plant species or communities.

4- The RF algorithm provides a reliable prediction of accuracy by using the out-of-bag samples. This could provide a tremendous saving of time and cost in data collection and analysis in remote sensing applications compared to the independent accuracy assessments method.

Overall, the use of hyperspectral indices and the RF algorithm for variables selection and classification techniques in this study proved a valuable tool to improve spectral discrimination between wetland plant species. However, the methods applied in Chapter one and two which were developed from fine spectral resolution (ASD) can be made operational by investigating their capability to discriminate between papyrus and its co-existing species using relatively coarser spectral resolution data such as AISA eagle spectra. Future research could also investigate the biochemical and biophysical variables that affect the canopy reflectance of the species studied.

Acknowledgments

We thank the Greater St Lucia Wetland Park authorities for allowing us access to the study areas. The contributions of Khalid Manssour in the field work are gratefully acknowledged. We extend our gratitude to Rick Lawrence for his thoughtful comments on facilitating the task. Our thanks go to Susan Davies for her help in proofreading. We thank the University of KwaZulu-Natal for providing the optimal conditions for the success of the study and assistance with the ASD. We thank Elfashir University for its contribution in funding this study.

CHAPTER SIX

Classifying papyrus vegetation (*Cyperus papyrus L.*) and its co-existing species using hyperspectral imagery and the random forest algorithm

This chapter is based on:

Adam, E., and Mutanga, O., (2010). Hyperspectral remote sensing of papyrus swamps. The 8th Conference of the African Association of Remote Sensing for the Environment (AARSE 2010). 25-29 October 2010. Addis-Ababa, Ethiopia.

.

Abstract

Mapping wetland plant species using multispectral remote sensing is challenging because of the small and mixed vegetation units in a wetland. The objective of this study was to examine the potential of airborne hyperspectral imagery to classify papyrus and its co-existing species in swamp wetland in St Lucia Park- South Africa. Hyperspectral image in 273 visible and near-infrared wavelengths (from 398 nm to 900 nm) and 2 m spatial resolution were acquired over the Dukuduku area by an Airborne Imaging Spectrometer for Applications (AISA) Eagle system. The canopy features of the papyrus and its co-existing species were identified using ground points and pixel-based average spectral reflectance at each wavelength from the acquired image, which was then used to develop a classification model. The RF classifier was used to classify the imagery using the randomForest package in R statistical program. The key wavelength determined by the integrated methods involved the RF and forward variable selection proposed in this study, and this could provide reasonable classification accuracy. Overall accuracy was 80.83 %, with class accuracies ranging from 86.67 % to 100 % and a kappa statistic of 0.74. The results also indicate that a subset of narrow band vegetation indices calculated from wavelengths allocated at the visible and red-edge portion of the spectrum could better improve the overall accuracy to 88.98 % and the kappa statistic to 0.85. The methods proposed in this study show considerable promise in mapping wetland vegetation at species level which is valuable for effective management of wetland ecosystems.

Keywords: Hyperspectral imagery. Variable selection. Vegetation indices. Random forest. Papyrus vegetation.

6.1 Introduction

Cyperus papyrus L., commonly called papyrus, belongs to the family Cyperaceae and is one of the most primary productive wetland plant species in eastern and central tropical Africa (Kyambadde *et al.*, 2004; Mnaya *et al.*, 2007). The natural distribution of papyrus swamps are thought to be confined to a belt across equatorial central Africa within 17^{o} N and 29^{o} S **(**Mnaya *et al.*, 2007**)**. The Greater St Lucia Wetland Park is within this belt and is one of the areas in which most extensive papyrus wetlands and swamps are found in South Africa (Adam and Mutanga, 2009). Papyrus commonly grows at the wetland edge anchored to the substratum, or sometimes creates extensive rafts of floating rhizomes in the middle of the wetland and at the lake-wetland interface (Kansiime *et al.*, 2005). In these wetland areas, papyrus forms a distinctive habitat type that supports a suite of specialist bird species and wildlife (Owino and Ryan, 2007; Grenfell *et al.*, 2009). Papyrus also plays a vital role in intercepting the materials moving from catchments to open water (Azza *et al.*, 2000; Serag, 2003; Kyambadde *et al.*, 2004). Moreover, promising results have been obtained in using wetland species, such as papyrus, as an alternative source of fuel in many countries in central Africa, such as Rwanda (Jones, 1983b; Muthuri and Kinyamario, 1989).

In most wetland habitats worldwide, human encroachment, intensified agricultural activities, and hydrological changes from construction of ditches, roads, and bridges in many parts of Africa have threatened the existence of papyrus (Mafabi, 2000; Maclean *et al.*, 2006; Owino and Ryan, 2007). As a result, this continued degradation in papyrus swamps represents a significant threat to biodiversity conservation (Owino and Ryan, 2007) and an increase in the sedimentation rates in the wetland areas (Grenfell *et al.*, 2009). Therefore, there is a need for accurate and quick field-wide monitoring for such an important plant species that could assist in making decisions to initiate protection and restoration programmes in the right place and at the right time (He *et al.*, 2005).

Mapping and monitoring wetland vegetation with traditional survey methods, such as hand mapping and Global Positioning Systems (GPS) receiver mapping, have proven to be highly accurate for small management areas (Cooksey and Sheley, 1997). However, these methods require intensive fieldwork ,including taxonomical information, collateral and ancillary data analysis, and the visual estimation of percentage cover for each species, which might be economically, technically, and logistically inadequate for wetland environments because of their high diversity and poor accessibility (Xie *et al.*, 2008; Adam *et al.*, 2009; Zomer *et al.*, 2009). Methods that take advantage of remote sensing can provide a time- and cost-effective technique to map and monitor such complex environments.

However, mapping wetland vegetation at species level using traditional remote sensing is challenging because of the lack of spectral resolution (1 to 7 bands), which limits the ability to map plant types based on the reflection and absorption of light at these few bands (Adam *et al.*, 2009). Furthermore, discrete wetland vegetation patches are usually smaller than the pixel size in most current spatial resolutions of multispectral images (Artigas and Yang, 2005; Zomer *et al.*, 2009). Therefore, with multispectral images the majority of pixels are a mixture of several plant species in various proportions even at high spatial scales (Zomer *et al.*, 2009). Hyperspectral sensors, on the other hand, enable the capturing of spectral data in many narrow bands (<10 nm) in up to 200 or more contiguous wavebands across the ultraviolet, visible, and infrared regions of the electromagnetic spectrum (Lillesand and Kiefer, 2001). Images from these new sensors, such as AISA and HYMAP, permit application of more complex spectral analyses and spectral unmixing techniques for a better separation of wetland vegetation at species level based on their unique light reflectance and absorption characteristics which can be especially useful for mapping percentage cover of the plant species (Artigas and Yang, 2005; Belluco *et al.*, 2006; Wang *et al.*, 2007; Adam *et al.*, 2009; Zomer *et al.*, 2009).

Previous attempts in classifying papyrus (*Cyperus papyrus L.)* using hyperspectral data include those by Adam and Mutanga (2009) who were able to implement a hierarchical method which used one-way analysis of variance ANOVA, classification and regression tree (CART), and distance analysis using hand-held spectroradiometer data to discriminate papyrus from its co-existing species (binary class) in the Greater St Lucia Wetland Park - South Africa. Another attempt in discriminating papyrus was that by Adam *et al.* (2009) They used the RF algorithm and field spectrometry data resampled to HYMAP resolution to discriminate papyrus and its co-existing species (multi-class classification). Their results indicated that there is a possibility of discriminating among papyrus and the other three species with an overall accuracy of 90.5 %. This overall accuracy can be improved to 96.5 % using vegetation indices calculated from field spectrometry data (Adam and Mutanga, In review).

The limitation of the above-mentioned studies is that the current operational airborne and spaceborne sensors, such as AISA and HYMAP, lack fine spectral resolution of the hand-held spectroradiometer which has a spectral range from 350˜ nm to 2500 nm (Mutanga, 2005). Therefore, it was recommended that the techniques implemented using a hand-held spectroradiometer should receive considerable additional testing with other airborne or spaceborne data (Adam *et al.*, 2009).

One of the most notable difficulties in hyperspectral data processing is the large data redundancy due to the strong correlation between wavebands that are adjacent (Shen, 2007). This high dimensionality requires sufficient training samples (Borges *et al.*, 2007) and computational processing which might be time-consuming and prohibitive in cost (Bajcsy and Groves, 2004). Therefore, techniques that reduce the high dimensionality without sacrificing significant information are highly sought after and feature selection or extraction tasks are often considered to be a practical and vital method in hyperspectral data processing and analysis (Borges *et al.*, 2007).

RF algorithm (RF), first developed by Breiman (2001), has recently been used as a classification and feature selection method to reduce the redundancy in hyperspectral data (Chan and Paelinckx, 2008; Adam *et al.*, 2009; Ismail, 2009). Random forest is a machine learning algorithm that employs a bagging (bootstrap aggregation) operation where a number of trees (*ntree*) are constructed based on a random subset of samples derived from the training data. Each tree is independently grown to maximum size based on a bootstrap sample from the training data set without any pruning, and each node is split using the best among a subset of input variables (*mtry*) (Breiman, 2001). The multiple classification trees then vote by plurality on the correct classification (Breiman, 2001; Lawrence *et al.*, 2006). The ensemble classifies the data that are not in the trees (out-of-bag or OOB data) and by averaging the OOB error rates from all trees, the random forest algorithm gives an error rate called the OOB classification error for each input variable (Breiman, 2001). Therefore, as part of the classification process, the RF algorithm produces a measure of importance of each input variable by comparing how much the OOB error increases when a variable is removed, whilst all others are left unchanged (Archer and Kimes, 2008). Studies have shown that the RF algorithm can be successfully used in hyperspectral data for feature selection as well as for classification purposes (Chan and Paelinckx, 2008; Adam *et al.*, 2009; Ismail, 2009) However, one of the shortcomings of the RF algorithm in selecting variables from very fine spectral resolutions such as spectroscopic data is that the selected relevant variables might still be auto-correlated (Strobl *et al.*, 2007).

The present study intends to examine the ability of hyperspectral imagery and the RF algorithm to discriminate amongst papyrus and its co-existing species in the Greater St Lucia wetland. More specifically, the objectives of the study were to: 1. examine the utility of the RF wrapper based approach for selecting the optimal number of hyperspectral wavebands in a multi-class application, 2. examine if the RF algorithm can accurately classify papyrus and its co-existing species in complex environments using hyperspectral airborne imagery, and 3. examine further whether vegetation indices calculated from hyperspectral imagery can improve the species classification using the RF algorithm.

6.2 Material and methods

6.2.1 Image acquisition and pre-processing

An Airborne Imaging Spectrometer for Applications (AISA) Eagle sensor was used to acquire hyperspectral images over a section of the study area (the Dukuduku forest and Futululu forest) in February 2009. The images were collected with 2 m spatial resolution, 272 wavebands (393 nm – 994 nm), and 2.04 nm to 2.29 nm spectral resolution. Images were taken at an altitude of approximately 1000 m above ground during cloudless periods in the daytime.

The image was atmospherically corrected using vicarious calibration techniques. Field spectral data of spectrally invariant targets (water body, tarred road surface) were collected during the flight campaign using an ASD spectrometer (Analytical Spectra Device). The field spectrometer senses in the range between 350 nm and 2500 nm incorporating the visible, near-infrared and short wave infrared bands. The field spectra were spectrally resampled to the spectral configuration of the AISA sensor and used to convert the AISA radiance data to absolute reflectance using the empirical line correction tool in ENVI software. A second order Savitzky-Golay function was used to smooth the AISA image as it presented some noise. A seven-band window size was used for the smoothing.

6.2.2 Field data collection

In order to achieve an accurate reference area for the classifier training, fieldwork was carried out concurrently with remote sensing campaigns to collect ground reference polygons of papyrus (*Cyperus papyrus L.*) and three co-existing species (*Phragmites australis*, *Echinochloa* *pyramidalis,* and *Thelypteris interrupta*) on February 2009. Leica Geosystems GS20 GPS Sensor with multiple-bounce filtering and post-differential correction was used to measure the position of the target species in swamp wetland with an accuracy of 0 m to 0.25 m after the post-processing differential correction. We randomly located transect lines within the study sites and sampled the target species (n = 4) randomly by drowning polygons along each transect by circumnavigating patches with an extent of 6 m to 8 m where the species present were more homogenous and unmixed. Ideally, a constraint on the size of the target species is that at least one entire AISA pixel (2 m × 2 m) should fall with each area covered by a homogenous species (Wang *et al.*, 2007). A point measurement of the central location of each polygon was also recorded. This method was rather difficult to implement because of the small vegetation species units with high spatial variability in a wetland environment (Adam *et al.*, 2009). However, this method resulted in 21 polygons for papyrus, 17 polygons for *Phragmites australis*, 14 polygons for *Echinochloa pyramidalis*, and 19 polygons for *Thelypteris interrupta* .These polygons were then used as reference data to generate regions of interest (ROIs). GPS field data were differentially corrected to enhance the accuracy using post-processing techniques.

The field data polygons (ROIs) were overlaid on the true colour composite AISA image to extract the pixels' spectra (6 m × 6 m) using ENVI software (ENVI, 2006). Only pixels that fell entirely within the measured polygons were included in the reference dataset, while the pixels that partially fell inside the polygons were discarded to avoid the problem of spectral mixing of the other plant species (Wang *et al.*, 2007). The reference values for each polygon were then averaged to represent one sample and used for development of models.

6.2.3 Selection of the optimal AISA spectral bands

Inter-band correlation exists in AISA imagery which provides redundant information. Reducing this high dimensionality in the spectral bands simplifies the model processing, decreases the running time of learning algorithm, and may improve the accuracy (Thenkabail *et al.*, 2004; Adam *et al.*, 2009). The RF algorithm and forward variables selection (FVS) were used to measure the importance of every AISA band in mapping the species and to select the optimal number of bands for better classification accuracy (Adam *et al.*, 2009). The RF algorithm developed by Breiman (2001) is a bagging (bootstrap aggregation) operation where multiple classification trees are constructed based on a random subset of samples derived from the training data. The optimization of the two parameters of RF includes the number of trees to be grown (*ntree*) and the number of variables to split the nodes of individual trees (*mtry*) that have firstly been optimized using the OOB estimates of error rate to guarantee high classification accuracy (Breiman, 2001; Adam *et al.*, 2009). The *ntree* values were tested from the default (500 trees) setting to 5500 tress with an interval of 1000 (Prasad *et al.*, 2006), while the *mtry* values were evaluated by creating RF ensembles for all possible *mtry* values (15). The setting that yielded the lowest OOB error was then used for any further analysis.

The importance of each AISA band (n =272) used in this study was calculated based on how much worse the classification accuracy (mean decrease in accuracy) would be if that variable (band) was permuted randomly using the internal out-of-bag estimates (Breiman, 2001; Lawrence *et al.*, 2006; Prasad *et al.*, 2006; Strobl *et al.*, 2007).The importance of each variable is estimated as follows: 1. the reflectance values of each wavelength is randomly permuted for the OOB samples, and then the modified OOB data are passed down each tree to get new predictions, 2. the difference between the misclassification rate for the modified and original OOB data over all the trees that are grown in the forest are then averaged, 3. this average is a measure of the importance of the variable and it is used as a ranking index which can be used to identify the wavelengths with relatively large importance in the classification process (Cutler *et al.*, 2007; Archer and Kimes, 2008; Chan and Paelinckx, 2008).

The FVS method was used to identify the optimal subset of wavelengths with the lowest misclassification error. The FVS method uses the ranking of wavelengths as determined by the RF algorithm. This method iteratively builds multiple random forests using the ranked wavelengths, and for each iteration two AISA bands were added to the model and the error was calculated using the OOB estimates of error. Initially, the top 2 ranked wavelengths are selected and for the next iteration, and then the top 4 ranked bands are selected (Adam *et al.*, 2009). This process was repeated for the maximum number of variables (bands) used in this study (n = 272).

6.2.4 Narrow band vegetation indices

Since remotely sensed measurements of vegetation canopies are affected by factors such as atmospheric absorptions, soil background and water, a normalization procedure using vegetation indices was also carried out in this study to minimize these influences ((Kokaly and Clark, 1999b; Mutanga and Skidmore, 2004a), and to enhance the possible difference in reflectance between the vegetation species (Qi *et al.*, 1995; Peña-Barragán *et al.*, 2006; Cho *et al.*, 2008; Adam and Mutanga, In review). Only five vegetation indices computed from field spectrometry data that yielded an overall accuracy of 96 % for discriminating papyrus and other species (Adam and Mutanga, In review) were adopted in this study. A full description of these vegetation indices is shown in Table 6.1. The RF algorithm was then used in order to evaluate the potential of these vegetation indices (n = 5) to discriminate papyrus and its co-existing species.

Table 6.1: Vegetation indices generated from AISA image and selected in this study

Vegetation indices	Abbreviation	Formula *	Reference
Normalized Difference Vegetation Index	NDVI	$\frac{R655 - R697}{R655 + R697}$	(Adam and Mutanga, In review)
Normalized Difference Vegetation Index	NDVI	$\frac{R655 - R705}{R655 + R705}$	(Adam and Mutanga, In review)
Plant Senescence Reflectance Index	PSRI	$\frac{R680 - R500}{R750}$	(Merzlyak *et al.*, 1999)
Blue/Green Index	BGI 1	$\frac{R400}{R550}$	(Zarco-Tejada *et al.*, 2005)
Pigment Index 4	PI 4	$\frac{R440}{R740}$	(Lichtenthaler *et al.*, 1996b)

* R = reflectance

6.2.5 Image classification

The *RandomForest* package in R software was used to classify the imagery (Liaw and Wiener, 2002). The bands that yielded the lowest OOB error using FVS were used as input variables in the RF model to classify the species. After optimizing the two parameters (*ntree* and *mtry*) of RF, the model was developed with 6500 classification trees (*ntree*) and with the default setting of the number of the bands to be split at each tree node (*mtry*). The RF model was developed using the entire reference data set, and accuracy was evaluated using the internally generated out-of-bag estimates of error. The out-of-bag estimates of error were developed using the one- third portion of the reference data set that was randomly excluded from development of each of the 6500 classification trees (Breiman, 2001; Lawrence *et al.*, 2006). Since OOB data are not used in the construction on any of the classification trees, it therefore is considered to be a type of cross-validation to estimate the prediction performance of the RF classifier (Breiman, 2001). The OOB accuracy assessment has been shown to be reliable and stable (Lawrence *et al.*, 2006; Prinzie and Van den Poel, 2008; Adam and Mutanga, In review). The OOB estimate of error was evaluated based on correctly classified pixels, and the confusion matrix was subsequently constructed to compare the true reference pixel with the pixels assigned by the classifier and to calculate the overall accuracy as well as the user and producer accuracy. Furthermore, a discrete multivariate technique called kappa statistics that uses the k (KHAT) statistic was also calculated to determine if one error matrix is significantly different from another (Cohen, 1960; Mutanga, 2005).

6.3 Results

6.3.1 Optimization of the random forest algorithm

The results of the RF parameters (*ntree* and *mtry*) optimization are shown in Figure 6-1. The results show that the OOB error rate is decreased substantially and becomes more stable as trees are added to the model. The optimum *mtry* value was found to be the default setting (n = 15) that was suggested by Liaw and Weiner (2002). The model yielded the lowest OOB error rate of 25.5 % with the default *mtry* (n = 15) and the high *ntree* (3500 to 9500). Therefore the default *mtry* and 3500 *ntree* were applied for all further analyses. Overall, the results clearly indicated that changes in RF parameters (*ntree* and *mtry*) influence the classification accuracy.

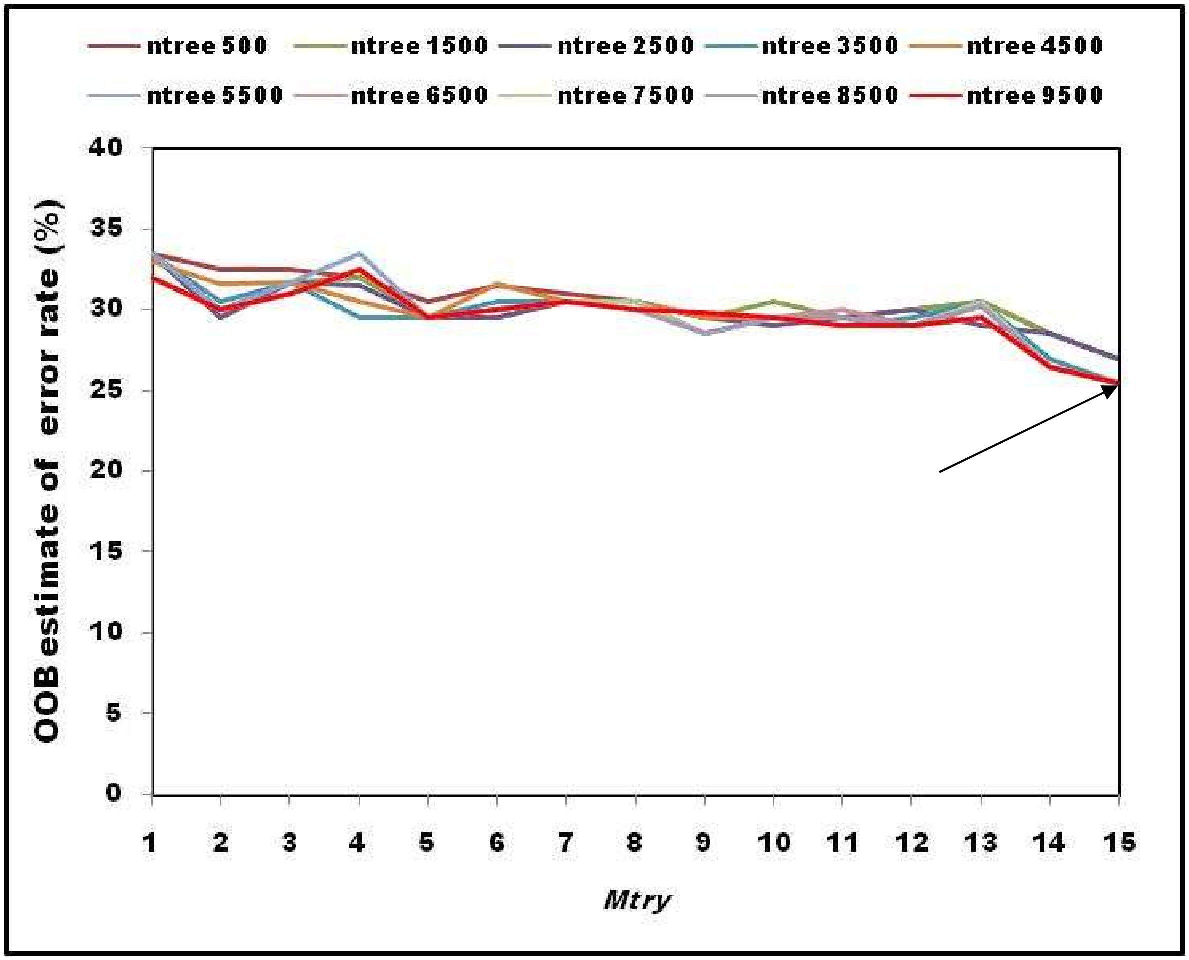

Figure 6.1. Optimizing random forest parameters (*ntree* and *mtry*) using the OOB estimate of error rate. The black arrow shows the optimal *mtry* number at the definite *ntree* number.

6.3.2 Variables selection

All AISA bands (n = 272) were included as potential variables for the RF model which was developed using 3500 *ntree* and a default setting of *mtry* (15). The entire model yielded an OOB error rate of 25.5 %. The importance of every single band of AISA imagery in mapping papyrus vegetation and the other species was calculated using the OOB estimate of error rate (Figure 6.2). The OOB error rate clearly showed the importance of each band based on how much the decrease in the classification accuracy would be if the data for that band were permuted randomly. Therefore, the high decrease in the accuracy means high importance and performance of the variable in mapping the target species.

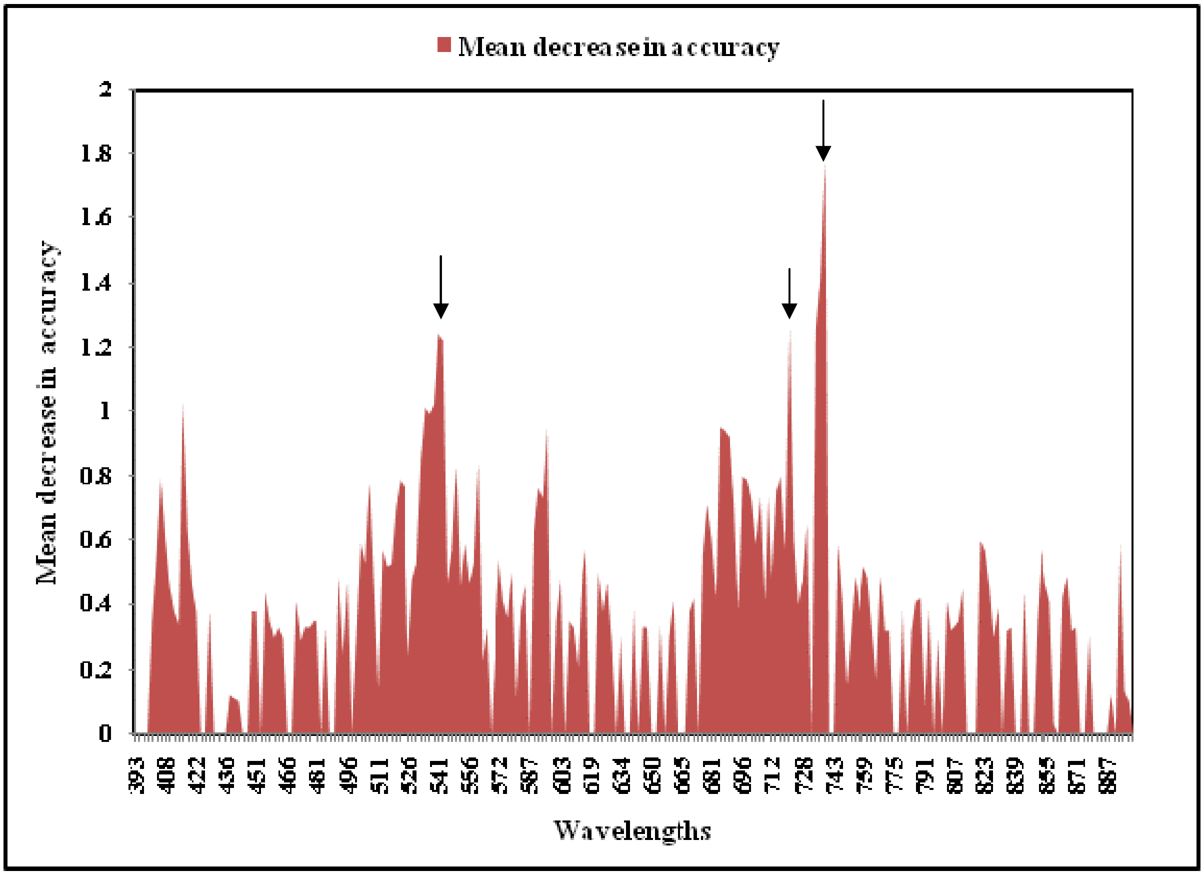

Figure 6.2. The importance of AISA bands in mapping papyrus and its co-existing species as determined by the RF model that yielded 25.5 % of OOB error rate. The black arrows show some of the most important bands.

Figure 6.2 clearly indicates that the most important bands are located in the green and red region (e.g. 541nm, 543 nm, 416 nm, 539 nm, 535nm, and 537 nm) and the red-edge portion of the spectrum (680 nm to 740 nm). The bands that show the highest importance in mapping papyrus and the other species are at 739 nm, 737 nm, 721 nm, 734 nm, and 541 nm (the ranking is based on the importance measures).

All the bands were then ranked according to their importance in mapping papyrus and the other species, and a forward variable selection was implemented in the top 100 bands which yielded an OOB error of 28 % for selecting the optimal number of bands as shown in Figure 6.3.

Results of forward variable selection (Figure 6.3) show that a subset including 8 bands located at 739 nm, 737 nm, 721 nm, 734 nm, 541 nm, 543 nm, 416 nm, and 539 nm resulted in the lowest OOB error rate of 19.17 % (misclassification rate) compared to 25.5 % when all bands (n

= 272) were used. The top 8 bands were then used as input variables in the final RF model to map papyrus and its co-existing species.

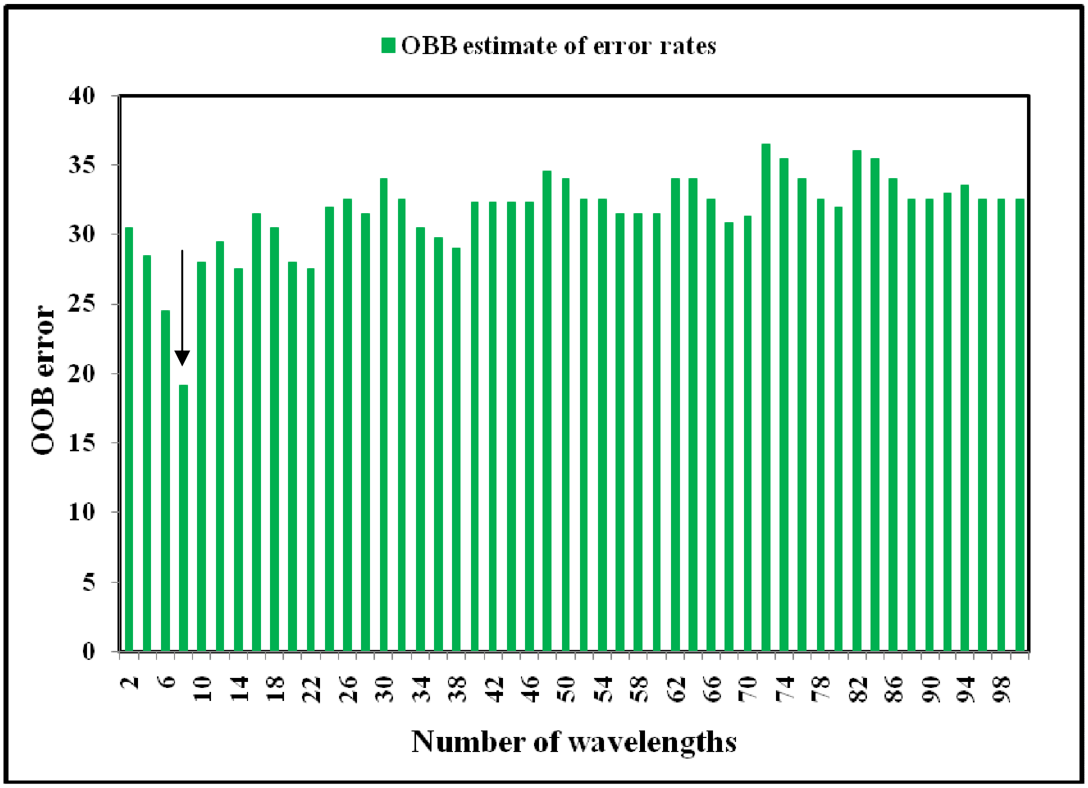

Figure 6.3: Selection of optimal number of variables (bands) using the forward variable selection method. The arrow shows the minimum number of bands that resulted in the lowest OOB error rate.

6.3.3 Classification and accuracy assessment

6.3.3.1 Using selected AISA raw bands

The 8 bands (739 nm, 737 nm, 721 nm, 734 nm, 541 nm, 543 nm, 416 nm, and 539 nm) were retained to classify the papyrus and the other species using the RF algorithm. The results indicate that the overall OOB error rate for all classes (CP *vs* PA, CP *vs* EP, CP *vs* TI, PA *vs* EP, PA *vs* TI, and EP *vs* TI) was 19.17 %. The confusion matrix in Table 6.2 clearly indicates that we could classify the papyrus vegetation and its co-existing species (n = 3) with an overall accuracy of 80.83 %.

Table 6.2: Testing the discriminatory performance of the RF classifier using the selected bands (n = 8) and the OOB method for estimating the error rate. The confusion matrix includes the overall accuracy, kappa statistic, user accuracy, and producer accuracy for *Cyperus papyrus* (CP), *Echinochloa pyramidalis* (EP), *Phragmites australis* (PA), and *Thelypteris interrupta* (TI)

Classes	CP	EP	PA	IT	Row total
CP	24	2	4	0	30
EP	4	22	4	0	30
PA	2	2	26	0	30
IT	2	3	0	25	30
Column total	32	29	34	25	120
Producer accuracy	= 75.86 %		Overall accuracy	= 80.83 %	
User accuracy	= 73.33 %		Kappa	= 0.74	

The high overall classification accuracy of 80.83 % and overall kappa statistic value of 0.74 achieved indicates the good performance of the variables selection method that was implemented in this study which was able to improve the overall accuracy using all 272 bands that yielded an overall accuracy of 74.5 %.

With respect to the class pairs accuracies, the selected bands (n = 8) yielded producer's accuracy that varied from 86. 67 % to 92.59 % and user's accuracy that varied from 84.62 % to 100 % for the three class pairs involving *Cyperus papyrus* (CP *vs* EP, CP *vs* PA, and CP *vs* TI). The lowest producer's accuracy and user's accuracy achieved were those that involved *Echinochloa pyramidalis* and *Phragmites australis*, and *Cyperus papyrus* and *Echinochloa pyramidalis* (86.67 % and 84.62 a respectively), while the highest user's accuracy was for the class pair that involved *Cyperus papyrus and Thelypteris interrupta (*100%) Table 6.3.

Table 6.3: Class pairs accuracies using the selected band (n =8) for *Cyperus papyrus* (CP), *Echinochloa pyramidalis* (EP), *Phragmites australis* (PA), and *Thelypteris interrupta* (TI)

Class pairs	Producer's accuracy	User's accuracy	Overall accuracy	Kappa
CP *vs* EP	91.67	84.62	88.46	0.77
CP *vs* PA	86.67	92.86	89.29	0.79
CP *vs* TI	92.59	100.00	96.08	0.92
EP *vs* PA	86.67	92.86	88.89	0.78
EP *vs* TI	100.00	89.29	94.00	0.88
TI *vs* PA	100.00	100.00	100.00	1.0

6.3.3.2 Using narrow band vegetation indices

Estimated classification accuracy of the species from out-of-bag data for the vegetation indices was 88.98 with a kappa statistic value of 0.85 (Table 6.4). As expected, the overall accuracy and kappa value were increased by 8.15 % and 0.11respectively compared with the use of raw bands. Producer's accuracy and user's accuracy were also increased by 10.81 % and 16.33 % respectively.

Table 6.4: Testing the discriminatory performance of the RF classifier using the selected vegetation indices (n = 5) and the OOB method for estimating the error rate. The confusion matrix includes the overall accuracy, kappa statistic, user accuracy, and producer accuracy for *Cyperus papyrus* (CP), *Echinochloa pyramidalis* (EP), *Phragmites australis* (PA), and *Thelypteris interrupta* (TI)

Classes	CP	EP	PA	IT	Row total
CP	24	2	4	0	30
EP	3	26	0	0	29
PA	2	0	27	0	29
IT	0	2	0	28	30
Column total	29	30	31	28	118
Producer accuracy	= 86.67 %			Overall accuracy	= 88.98 %
User accuracy	= 89.66 %			Kappa	= 0.85

Producer's accuracy and user's accuracy, which are more meaningful for the individual classes, are shown in Table 6.5. The results presented in Table 6.5 show the feasibility of using the vegetation indices in the designation of the RF classification algorithm, having improved the producer's and user's accuracy with the range of 0.43 % to 13.33 % and 0.24 % to 7.14 % respectively for most of the class pairs in comparison to raw bands spectral classifications. For comparison, the producer's and user's accuracy also for the raw bands and vegetation indices are presented in Figure 6.4.

Table 6.5: Class pairs accuracies using the selected vegetation indices (n = 5) for *Cyperus papyrus* (CP), *Echinochloa pyramidalis* (EP), *Phragmites australis* (PA), and *Thelypteris interrupta* (TI)

Class pairs	Producer's accuracy %	User's accuracy %	Overall accuracy %	Kappa
CP *vs* EP	92.86	89.66	90.91	0.82
CP *vs* PA	87.10	93.10	89.47	0.79
CP *vs* TI	92.31	100.00	92.86	0.86
EP *vs* PA	100.00	100.00	100.00	1.00
EP *vs* TI	100.00	93.33	96.43	0.93
TI *vs* PA	100.00	100.00	100.00	1.00

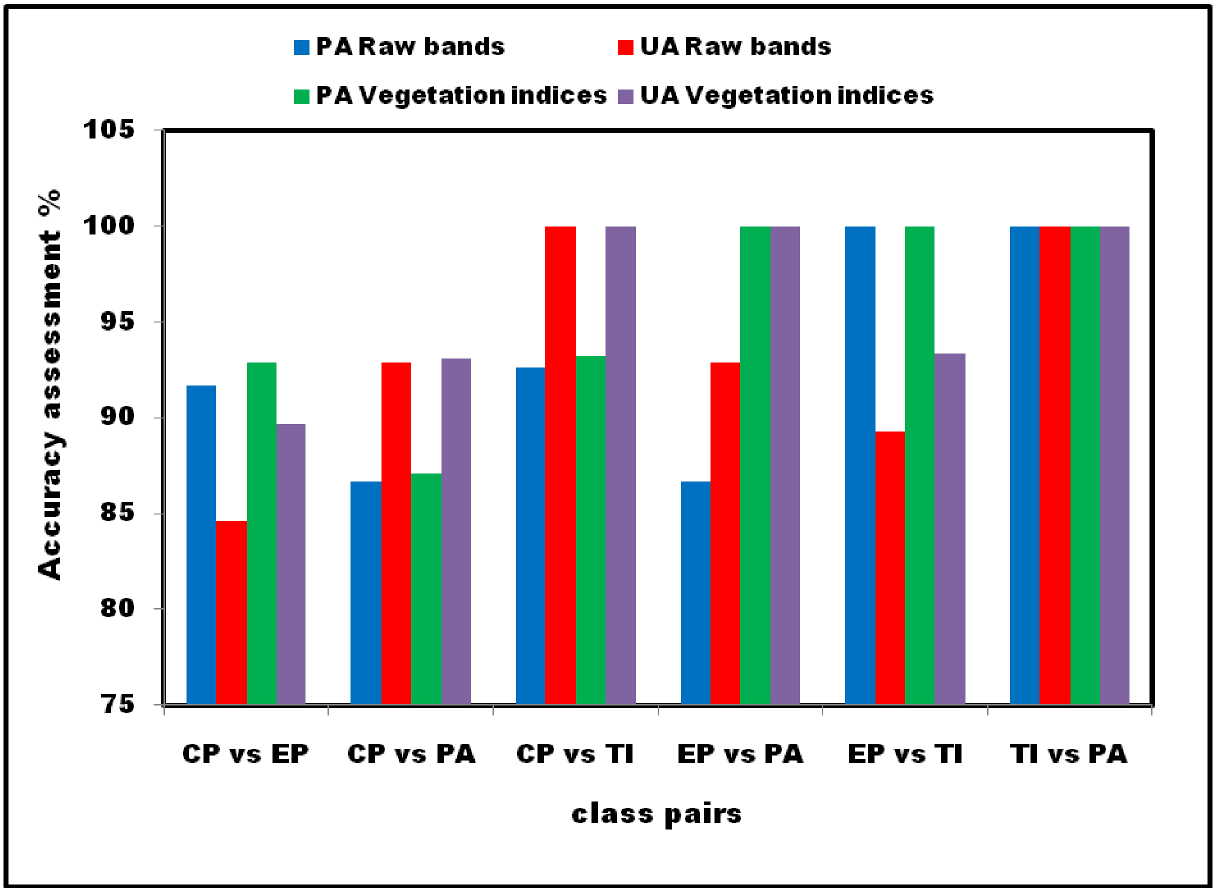

Figure 6.4: Producer's accuracy (PA) and user's accuracy (UA) generated from the use of AISA raw bands (n = 8) and vegetation indices (n = 5) for each class pair of the species *Cyperus papyrus* (CP), *Echinochloa pyramidalis* (EP), *Phragmites australis* (PA), and *Thelypteris interrupta* (TI))

6.4 Discussion

Effective management of wetland vegetation species requires accurate knowledge of their spatial distribution and density to assist in the effort to protect and sustain this valuable ecosystem. This can be achieved to different degrees by the processing of different remotely sensed data. This study attempted to scale up the method proposed by Adam *et al*. (2009) to an airborne sensor for mapping papyrus and its co-existing species. We tested the utility of the AISA imagery with a spectral resolution of 272 visible and near-infrared (NIR) wavebands and a spatial resolution of 2 m to map papyrus and its three co-existing species in a swamp wetland in Greater St Lucia Wetlands Park –South Africa. Our results demonstrated that papyrus vegetation and its co-existing species can be separated from each other with a high level of overall accuracy (80.83 %).

The study emphasized the main obstacle in classifying and characterizing the distribution of papyrus vegetation and its associated plant species. This included collecting sufficiently accurate and enough ground truth points for training data in the image, since spatial variation and diversity in the wetland vegetation is very high, and not easily accessible (Bajjouk *et al.*, 1998; Adam *et al.*, 2009; Artigas and Pechmann, 2010). This obstacle could result negatively on the classification accuracy, since a shift of one pixel may induce a significant error, and therefore the overall results will not be reliable (Artigas and Pechmann, 2010). We believe, however, that this problem was overcome by ensuring the selection of ground reference area (ROIs) that contain a single species over more than 80 % of the area and are larger than the pixel size of AISA imagery (2 m). Moreover, the boundaries of the ROIs were accurately delimited using differential GPS with a minimum accuracy of ± 1 cm (Belluco *et al.*, 2006).

Results from this study show that 8 bands of AISA selected as the optimal number using the separability statistics method developed in this study yielded classification accuracies that were better than those obtained when the entire hyperspectral bands (n = 272) were put into the RF classifier algorithm. This is beneficial for cost-effective wetland vegetation mapping in terms of reducing the time and space needed to process and store the hyperspectral data. Moreover, the variables selection method used in this study which integrated RF and FVS allowed direct measuring of the importance of variables (bands) at the same time as the classification process of hyperspectral data which is recommended in remote sensing techniques (Guyon and Elisseeff, 2003; Granitto *et al.*, 2006; Adam *et al.*, 2009; Ismail, 2009). We believe that our remarkable results in this regard show the usefulness of the RF algorithm as a technique for reducing the dimensionality of hyperspectral data. It is, therefore, worth considering RF as a useful technique for variables selection in hyperspectral remote sensing in the future.

Among the bands selected in this study, many are located in the red-edge portion of the spectrum (739 nm, 737 nm, 721 nm, and 734 nm). These bands are within ± 12 nm from known bands that are selected for discriminating the same species in other studies. These are 710 nm and 725 nm (Adam *et al.*, 2009) 745 nm, and 746 nm (Adam and Mutanga, 2009). Other studies have also reported the usefulness of the red-edge portion for mapping wetland vegetation (Daughtry and Walthall, 1998; Thenkabail *et al.*, 2002; Vaiphasa *et al.*, 2005). The red-edge portion has been found to be sensitive to chlorophyll and biomass variation (Sims and Gamon, 2002; Mutanga and Skidmore, 2007). The rest of the bands selected in this study are located in the visible region of the spectrum (541 nm, 543 nm, 416 nm, and 539 nm) which are ± 12 nm from the known visible bands selected for mapping wetland vegetation in previous studies such as 550 nm (Daughtry and Walthall, 1998; Thenkabail *et al.*, 2002) and 404 nm (Schmidt and Skidmore, 2003). According to Tucker (1977), the variations in the vegetation spectra reflectance in the visible region are primarily determined by the concentration of chlorophylls and carotenoids.

As we expected, the RF classifier adopted in this study produced high classification accuracies (85%). The RF classier has also been recently applied successfully in the classification of hyperspectral remote sensing data, and overall accuracies of more than 80 % have also been reported (Gislason *et al.*, 2006; Lawrence *et al.*, 2006; Adam *et al.*, 2009). The method has many advantages such as that it is not sensitive to the noise or overtraining and only two user defined parameters are needed. Therefore, the RF classifier could be considered to be a very desirable method for classification of hyperspectral remote sensing data (Lawrence *et al.*, 2006). The reliability of the internal method of the OOB estimate of accuracy (Gislason *et al.*, 2006; Lawrence *et al.*, 2006) that was adopted in this study has provided a tremendous saving of time and labour to collect separate accuracy assessment data which was difficult under the conditions of the study sites located in the swamp wetland.

The overall classification accuracy (80 %) we achieved in this study using AISA bands is 9.5 % lower than that which has been reported by Adam *et al*. (2009). This can be explained by the fact that the AISA airborne sensor used in this study lacks the fine spectral resolution of the ASD.

To further increase the classification accuracy of the species, we tested the utility of a subset of vegetation indices (Adam and Mutanga, In review). These vegetation indices include Plant Senescence Reflectance Index, Blue/Green Index1, Pigment Index4, and the modified NDVIs involving a combination of a narrow band in the red portion (655 nm) with two wavelengths in the red-edge position (697 nm and 705 nm). An interesting result from this study is the finding that these vegetation indices were able to increase the overall classification accuracy of papyrus and its co-existing species from 80.83.5% to 88.98 % and to increase the overall kappa statistic from 0.74 to 0.85 (Table 6.2 and 6.4). This result is identical to the finding of Adam and Mutanga (in review) who reported that narrow band vegetation indices preformed better than the selected raw bands in discriminating among papyrus and its co-existing species using field spectrometery data. This better performance of the vegetation indices could possibly be explained by the fact that vegetation indices enhance the possible difference in reflectance between the vegetation species (Qi *et al.*, 1995; Peña-Barragán *et al.*, 2006; Cho *et al.*, 2008) and minimize the influences of atmospheric absorptions, soil background, and water on vegetation canopies (Kokaly and Clark, 1999b; Mutanga and Skidmore, 2004a).Since leaf biochemical and biophysical features were not measured in this study, there is therefore the need to explain why these vegetation indices showed a higher accuracy in discriminating papyrus and its co-existing species.

6.5 Conclusions

The RF classifier was appropriate for this study because it does not require a separate accuracy assessment data set. This was useful because there was a tremendous saving of time and labour in collecting more ground truth points in such swamp areas, which are not easily accessible. We caution, however, that because RF is a supervised classification technique for working with areas of mixed vegetation species unbiased sampling and careful fieldwork is necessary for acquiring accurate information of training samples in order to get reliable classification results.

The results from this study demonstrate that airborne hyperspectral imagery can be a useful source of data for distinguishing papyrus and its co-existing plant species. However, in order to better understand the spatial variations of papyrus quantity and quality, it would be useful if estimation of biophysical and biochemical parameters of papyrus such as biomass is also investigated in further studies.

In summary, the methods and procedures presented in this study can be used for mapping other wetland plant species. The RF algorithm applied to hyperspectral data was able to provide high accuracy in the classification model. It remains to be tested in regression model using hyperspectral data.

Acknowledgements

The Department of Science and technology (DST)-South Africa provided the hyperspectral imagery for this study. We extend our gratitude to Dr Moses Cho for the acquisition of the hyperspectral imagery. We also appreciate the generous help provided by Abd Allah Ibrahim and Khalid Manssour in the reflectance data collection and ground verification

CHAPTER SEVEN

Estimating papyrus (*Cyperus papyrus L.)* biomass using narrow band vegetation indices and the random forest regression algorithm

This chapter is based on:

Adam, E., Mutanga, O., and Ismail, R., (accepted). Estimating papyrus (*Cyperus papyrus*) biomass using narrow band vegetation indices and the random forest regression algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing.*

Abstract

Accurate estimates and mapping of wetland vegetation quality such as biomass have increasingly been identified as critical components for an efficient wetland monitoring and management system. Traditionally, biomass predictions are made using direct field measurement methods. These methods do not offer real-time data, and are inadequate for poorly accessible areas. Methods that take advantage of remote sensing can offer powerful techniques for predicting vegetation biomass. In this study, we investigated the use of vegetation indices derived from field spectrometry data to estimate papyrus (*Cyperus papyrus L.*) biomass. Papyrus characterizes most of the wetlands in tropical Africa. Spectral and above ground biomass measurements were collected at three different areas in the Greater St Lucia Wetland Park, South Africa. We evaluated the potential of narrow band normalized difference vegetation index (NDVI) calculated from all possible two-band combinations between 700 nm and 1000 nm. Subsequently, we utilized the RF (RF) algorithm as a modelling tool for predicting papyrus biomass. The results showed that papyrus biomass can be estimated at full canopy level under swamp wetland conditions (R^2 = 0.73; RMSEP = 276 g/m^2; 8.6 % of the mean). From our results, the RF algorithm has proved to be a robust feature selection method in identifying the minimum number (n = 4) of narrow band NDVIs that offered the best overall predictive accuracy. This lowest prediction error (RMSEP = 276 g/m^2; 8.6 % of the mean) was obtained using four NDVIs computed from bands at (740 nm and 853 nm), (741 nm and 853 nm), (741nm and 847 nm), and (749 nm and 776 nm). It was recommended that these promising results can be upscaled to spaceborne or airborne sensors such as HYMAP or Hyperion for predicting vegetation biomass in wetland areas using remotely sensed data.

Keywords: Above ground biomass. Field spectrometer measurements. NDVI. Random forest. Variables selection.

7.1 Introduction

Papyrus (*Cyperus papyrus L.*) is increasingly being recognized as the most biomass productive plant species in the tropical wetlands in Africa (Muthuri and Kinyamario, 1989). Papyrus plays a vital role in hosting habitats for wildlife species (Owino and Ryan, 2007), and it has a major influence on the grazing distribution patterns of livestock especially in dry seasons (Muthuri and Kinyamario, 1989). Furthermore, promising results have been obtained in using papyrus as an alternative source of fuel in many countries in central Africa such as Rwanda (Jones, 1983b; Muthuri and Kinyamario, 1989).

Despite the relative importance of papyrus, human encroachment and intensified agricultural activities in many parts of Africa have threatened the existence of papyrus functions (Mafabi, 2000; Maclean *et al.*, 2006; Owino and Ryan, 2007). The continued degradation in papyrus habitat represents a significant threat to biodiversity conservation particularly for papyrus-specialist birds and other papyrus-reliant species in many African countries (Maclean *et al.*, 2006; Owino and Ryan, 2007). Therefore, efficient techniques that can spatially and temporally monitor the stability of the productivity of papyrus ecosystems and whether significant changes are taking place in these swamp ecosystems are required. Such techniques require up-to-date spatial information on the distribution of papyrus vegetation. Also, the variation in the quality and quantity of papyrus vegetation is critical for a better understanding of the productivity and functioning of papyrus swamps (Adam and Mutanga, 2009). Previous studies have shown the possibility of discriminating papyrus from its co-existent species using hyperspectral remote sensing (Adam and Mutanga, 2009; Adam *et al.*, 2009). However, timely assessment and mapping of both papyrus species and above ground biomass (AGB) variation is needed to facilitate a better understanding of the species-quality interaction in their spatial distribution (Mutanga, 2004).

Traditional methods such as field measurements have been used to estimate papyrus AGB (Jones and Muthuri, 1997; Serag, 2003; Boar, 2006). However, these traditional methods require sufficient numbers of samples which is expensive, time-consuming, and difficult to implement, especially in large and inaccessible areas (Lu, 2006). Biomass estimation based on remote sensing has increasingly attracted scientific interest because of its cost-effectiveness and benefit of repetitively collecting digital data. Additionally, researchers have shown that there are high correlations between spectral bands and vegetation biomass (Lu, 2006). In this regard, broad band remote sensing has been widely used to model the spatial and temporal variability of vegetation biomass over large wetland areas (Ramsey and Jensen, 1996; Moreau *et al.*, 2003; Rendonga and Jiyuanb, 2004; Proisy *et al.*, 2007). The shortcoming of broad band satellite data is that the high spectral variation and shadows caused by canopy and topography may create difficulty in developing an accurate biomass estimation model that can differentiate between vegetation and the soil background (Lu, 2006; Numata *et al.*, 2008). Some studies have demonstrated that vegetation indices (VI) have the potential of overcoming some of these problems (Elvidge and Chen, 1995; Todd *et al.*, 1998). The most commonly used vegetation indices, which are sensitive to biophysical and biochemical variation in vegetation, are computed from the red and near-infrared (NIR) portions of the electromagnetic spectrum (Asrar, 1989; Cho *et al.*, 2007). These vegetation indices respond to the difference between the reflectance in the visible portion because of the chlorophyll absorption and high reflectance in the NIR due to the multiple scattering effects of vegetation (Elvidge and Chen, 1995).

The normalized difference vegetation index (NDVI) (Rouse *et al.*, 1973) has been widely used during the last decades for modelling the spatial variability of AGB based on broad band sensors (50 nm-100 nm) such as NOAA and Landsat Thematic Mapper (Moreau *et al.*, 2003; Lu *et al.*, 2004). However, the major limitation of NDVI is that the broad band NDVI uses average spectral information over a wide range of the spectrum which results in loss of critical information (Hansen and Schjoerring, 2003; Numata *et al.*, 2008). Furthermore, NDVI calculated from broad band sensors asymptotically approach a saturation level after a certain AGB (about 15 kgm^{-2}) or vegetation age (15 years in tropical forest) (Steininger, 2000; Lu and Batistella, 2005). Therefore, NDVI yields poor estimates during peak growing seasons and in more densely vegetated areas (Thenkabail *et al.*, 2000; Mutanga and Skidmore, 2004a). In general, the estimation of AGB is still a challenging task, especially in those study areas with mixed species, densely vegetated environments, and complicated canopy structure (Adam *et al.*, 2010). Given these limitations and challenges, there is a need to develop or to improve techniques for better estimation of AGB in highly diverse and densely vegetated areas such as wetlands where there is almost 100 % vegetation cover.

More recently, the development of field hyperspectral remote sensing has opened new perspectives for investigating the most powerful narrow bands to be used in VIs formulation and for maximizing the sensibility of VIs to AGB based on the whole electromagnetic spectrum (350 nm – 2500 nm) rather than focusing on the red and NIR bands (Hansen and Schjoerring, 2003; Mutanga and Skidmore, 2004a; Cho *et al.*, 2007; Fava *et al.*, 2009). The use of NDVI calculated from narrow bands has been found to be one possibility to reduce the data saturation problem (Mutanga and Skidmore, 2004a). However, some authors note that the strengths of the rich hyperspectral bands have not be exploited because only two bands from the red and near-infrared regions are used to formulate the indices (Cho *et al.*, 2007). Alternatively, multiple linear regression (MLR) methods based on more than two bands have been shown to be effective in estimating AGB (Lu, 2006). However, identifying suitable variables for developing a multiple regression model is often critical because some variables are weakly correlated with AGB or are highly correlated to each other (Lu, 2006). Given this problem, a powerful method for identifying the most useful narrow band indices to improve the prediction of AGB is essentially required (Lu, 2006).

Ensemble methods like RF (Breiman, 2001) have been used to enhance the prediction accuracy in the field of ecology (Prasad *et al.*, 2006; Grimm *et al.*, 2008). Results from these studies concluded that the RF algorithm and bagging have similar abilities for improving prediction accuracy, with slightly better performance by the RF. From the field of remote sensing, ensemble approaches have been widely applied in different fields as a classification algorithm (Ham *et al.*, 2005; Pal, 2005; Gislason *et al.*, 2006; Lawrence *et al.*, 2006; Adam *et al.*, 2009). To the best of our knowledge, only Ismail and Mutanga (2009) investigated the use of the RF algorithm in regression type applications for predicting *S. noctilio* induced water stress in *P. patula* trees using hyperspectral data. Results from the study showed that the RF algorithm outperformed bagging and boosting (R^2 = 0.73). Therefore, in this study we further investigated the performance of regression tree ensembles on variables selection and for predicting AGB of papyrus in a complex environment which has been overlooked in scientific research.

Thus, the research objectives were: (i) to evaluate the utility of narrow band NDVI derived from field spectrometry measurements for estimating papyrus AGB in complex and densely vegetated canopies, and (ii) to test the performance of the RF algorithm in a regression application (i.e. identifying the best hyperspectral indices and predicting AGB). To achieve these tasks, a field experiment was planned to collect AGB and spectral data from papyrus vegetation in the summer of 2009 at the Greater St Lucia Wetland Park, South Africa, which is characterized by a composition of mixed species. The vegetation indices (NDVI) were calculated, and the predictive performance of the regression tree ensembles was then determined using training or calibration and test data sets.

7.2 Material and methods

7.2.1 Field spectral measurements and biomass harvesting

Random sampling was adopted in this study. Hawth's Analysis tool was used to generate 50 random points on a land cover map developed from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) imagery. The sample points were subsequently uploaded into a GPS that was used to navigate to the field sites i.e. Futululu Park, and the Mfabeni and Mkuzi swamps. Once the sample site was located, a 30 m by 30 m vegetation plot was created to cover a homogenous area of the papyrus. Three subplots (1 m × 1 m) were then randomly selected within each plot to measure the spectral reflectance.

All the spectral measurements were collected in the summer of 2009 between 10:00 am and 02:00 pm under sunny and cloudless conditions using the Analytical Spectral Devices (ASD) FieldSpec® 3 spectrometer. The spectrometer measures wavelengths ranging from 350 nm to 2500 nm with a sampling interval of 1.4 nm for the 350 nm to1000 nm spectral region, and a 2.0 nm sampling interval for the 1000 nm to 2500 nm spectral region. The ASD has a spectral resolution of between 3 nm and 10 nm (ASD Analytical Spectral Devices Inc., 2005). A white reference spectralon calibration panel was used every 5 to 10 measurements to offset any change in the atmospheric condition and irradiance of the sun. Accompanying the field spectral measurements, metadata such as the sites' description (coordinates, altitude, and land cover class) and general weather conditions were also recorded (Milton *et al.*, 2009). From each subplot (1 m × 1 m) approximately 5 to 10 field spectrometer measurements were randomly taken at nadir from 1 m using a 5^o field of view. This resulted in a ground field of view of about 18 cm in diameter, which was large enough to cover a cluster of papyrus and reduce the background effects caused by soil and water (Mutanga *et al.*, 2004). These spectral measurements were then averaged to obtain the final spectral measurement for each vegetation plot (30 m × 30 m).

After spectral measurements, AGB was clipped within the subplots (1 m × 1 m). All dry material was removed from the clipped plants and fresh biomass was then measured immediately using a digital weighing scale. Average fresh AGB per plot was then calculated from the three subplot measurements (Cho and Skidmore, 2009).

7.2.2 Data analysis

7.2.2.1 Narrow band indices

The narrow band NDVI-based vegetation indices were computed in this study from all possible two-band combinations using all the red, red-edge, and NIR bands (i.e. 600 nm to 1000 nm). These indices and spectral regions were selected because they are the most commonly used in estimating biomass and crop yield (Thenkabail *et al.*, 2000; Mutanga and Skidmore, 2004a; Cho *et al.*, 2007). The discrete 401 narrow bands allowed a computation of N*N = 160,801 narrow band indices using the principle of the NDVI calculation as follows:

$$NDVI = \frac{R_{(i,n)} - R_{(j,n)}}{R_{(i,n)} + R_{(j,n)}}$$

Where $R_{(i,n)}$ and $R_{(j,n)}$ are the reflectance of any two bands from the selected bands for spectral sample (n).

7.2.2.2 Random forest regression ensemble

The RF algorithm (Breiman, 2001) was used in this study to predict the AGB of papyrus (g m^{-1}). The RF algorithm was developed to reduce the instability and the variance of a single regression tree. The algorithm generates multiple bootstrap samples from the original training data set with replacement to create multiple regression trees (*ntree*). The model allows these regression trees to grow to maximum size without pruning. Each tree is grown in RF with a randomized subset of predictors (*mtry*) to determine the best split at each node of the tree (Breiman, 2001). The results from each aggregation are then averaged to get the overall prediction accuracy. Because there is a large number of trees, RF achieves low bias and low variance (Grimm *et al.*, 2008).

When a bootstrap sample is drawn, about 37 % of the dataset is excluded from the sample and the remaining data are replicated to bring the dataset to full size. This dataset is defined as ‘in bag’ data, while the excluded dataset (approximately 37 %) is known as the ‘out-of-bag’ data (OOB) (Breiman, 1996). For each tree in the ensemble, the RF algorithm also calculates the mean square error as the difference between predictions (i.e. mean square error) made using the OOB data and the 'in bag' data, known as the OOB error (Prasad *et al.*, 2006; Palmer *et al.*, 2007). The OOB estimate of error is considered to be a reliable assessment of predictive accuracy since the OOB data were not used to build or prune any regression trees in the ensemble. The OOB error estimate is considered to be a form of cross-validation (Svetnik *et al.*, 2003) and provides a good and reliable internal estimate of error (Breiman, 1996, 2001; Prasad *et al.*, 2006; Grimm *et al.*, 2008; Ismail and Mutanga, 2009). Some researchers have suggested that it may not be necessary to have an independent validating dataset (Lawrence *et al.*, 2006).This is of particular interest regarding wetland areas, since data collection is difficult due to the poor accessibility of areas. Additionally, the OOB data allow for the evaluation of the importance of each variable in the prediction by determining how much the prediction error would increase if the OOB data of that variable were permuted (Prasad *et al.*, 2006).

In the RF algorithm there are two parameters which need to be optimized by the user: the number of trees (*ntree*) in the forest and the randomly selected number of variables tried at each node (*mtry*) (Breiman, 2001). The default value of *ntree* is 500, while the default *mtry* value for regression applications is one-third of the total number of predictors. In this study, the *ntree* values were tested from the default setting 500 to 5500 with an interval of 1000 (Prasad *et al.*, 2006), while the *mtry* was evaluated by creating RF ensembles for all possible *mtry* values (20) (Ismail and Mutanga, 2009). The optimal values of *ntree* and *mtry* were then selected based on the lowest root mean square error of calibration (RMSEC).

The *randomforest* library (Liaw and Wiener, 2002) developed in the R package for statistical analysis (R Development Core Team, 2007) was employed to implement the RF algorithm.

To validate the performance of the RF algorithm (Lawrence *et al.*, 2006), the data were randomly divided into 70 % training or calibration and 30 % test data samples (n = 32 and 14 respectively). Regression analyses were performed on the calibration dataset using the OOB estimates of error. The test data set was used to validate the predictive performance of the RF algorithm (Lawrence *et al.*, 2006; Ismail and Mutanga, 2009). A one-to-one relationship between measured and predicted biomass values was then established. The coefficients of determination (R^2) for calibration and prediction as well as RMSEC and root mean square error of prediction (RMSEP) values were reported.

7.2.2.3 Selection of the predictive variables

The narrow band indices NDVIs computed from all possible two-band combinations of 401 bands were ranked based on the correlation coefficient = r (R^2 = coefficient of determination). The top 20 NDVIs that yielded the highest R^2 were then selected for further analysis in order to simplify the modelling process (Mutanga and Skidmore, 2004a).

In order to simplify the modelling process, it was necessary to identify the smallest number of NDVIs that offered the best predictive performance for AGB. The RF procedure measured the importance of the top 20 band combinations from the training dataset (70 %) based on the mean decrease in accuracy. The variables were ranked according to their importance. We subsequently used this ranking to indentify the sequence in which to discard the least important variables (NDVI) using backward elimination function (Ismail and Mutanga, 2009). The backward variable selection process iteratively builds multiple random forests for regression. At each iteration, a new forest was developed after gradually eliminating one of the least promising narrow band NDVIs (n = 20), and RMSEC was calculated. We further evaluated the selection of the best subset using an independent test dataset (Kohavi and John, 1997). We compared the performance of OOB with both the hold out test dataset and the 10 fold cross-validation (Ismail and Mutanga, 2009). The nested subset of variables (NDVI) that yielded the lowest RMSEC was then selected as the optimal variable for biomass prediction.

7.3 Results

As a precursor to examining the relationship between AGB and NDVIs, descriptive statistics of biomass were generated and the results are shown in Table 7.1.

Table 7.1: Descriptive statistics of the measured above ground biomass.

	Sample No	Unit	Mean	S.D.	Minimum	Maximum	Range
Biomass	47	g/m^2	3221.362	562.3853	2367	4305	1938

7.3.1 Hyperspectral indices (NDVI) and biomass

The reflectance values of narrow band hyperspectral data contained 401 discrete channels located in the red or far- red and NIR (600 nm – 1000 nm) allowed the computation of 160,801 NDVIs for biomass estimation. Analysis of the correlation coefficients, R^2, between the entire possible two narrow band NDVIs (n = 160,801) and AGB of papyrus is shown in Figure 7.1. It can be clearly seen from this figure that there is a wide variation in strength of the relationship between NDVIs and AGB. The R^2 values range between 0.00 and 0.83. The band combinations involving the far-red-edge bands located from 720 nm to 850 nm range yielded the strongest correlations (0.73 to 0.83).

The NDVIs were then ranked based on their correlation coefficients, and the top 20 two-band combinations that yielded the highest R^2 values were then selected and recorded as shown in Table 7.2 for further analysis.

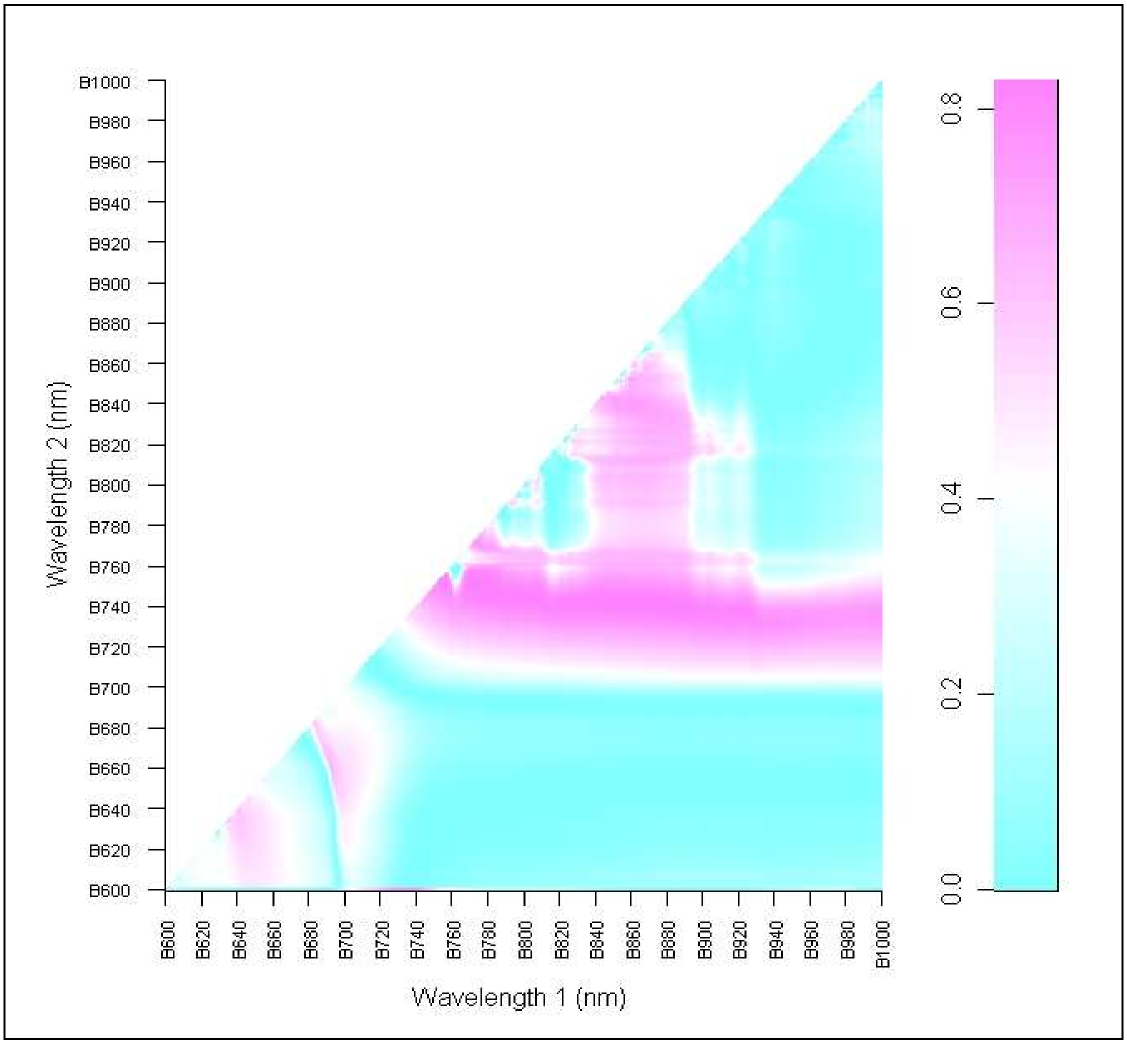

Figure 7.1. Contour plot representing the correlation coefficients (R^2) of the linear regression between above ground green biomass and NDVIs obtained from all possible two-band combinations using bands located from 600 nm to 1000 nm.

Table 7.2: The top 20 NDVIs that yielded the highest correlation coefficients for papyrus biomass

Rank	Wavelength 1 (nm)	Wavelength 2 (nm)	R	R^2
1	741	853	0.910	0.829
2	740	853	0.910	0.828
3	741	847	0.910	0.828
4	749	776	0.910	0.828
5	741	845	0.910	0.828
6	740	865	0.910	0.828
7	740	849	0.910	0.828
8	749	778	0.910	0.827
9	750	773	0.910	0.827
10	740	840	0.908	0.825
11	749	771	0.908	0.825
12	752	773	0.908	0.825
13	743	809	0.908	0.825
14	745	803	0.904	0.818
15	739	895	0.904	0.817
16	739	822	0.904	0.817
17	740	800	0.904	0.817
18	754	770	0.904	0.817
19	744	774	0.904	0.817
20	752	784	0.900	0.810

7.3.2 Parameters optimization of the random forest regression model

The results of optimizing RF parameters (*ntree* and *mtry*) are shown in Figure 7.2. Based on the lowest RMSEC, the default value of *mtry,* which is one-third of the total number of variables (in this study = 7), is often the best choice with different values of *ntree*. With respect to *ntree* values, the results show that the model performs better (low RMSEC) when the *ntree* value is high (*ntree* = 5500) (Figure 7.2). Overall, the results showed that changes in the parameters of the RF regression (*ntree* and *mtry*) affect the error of prediction of the model.

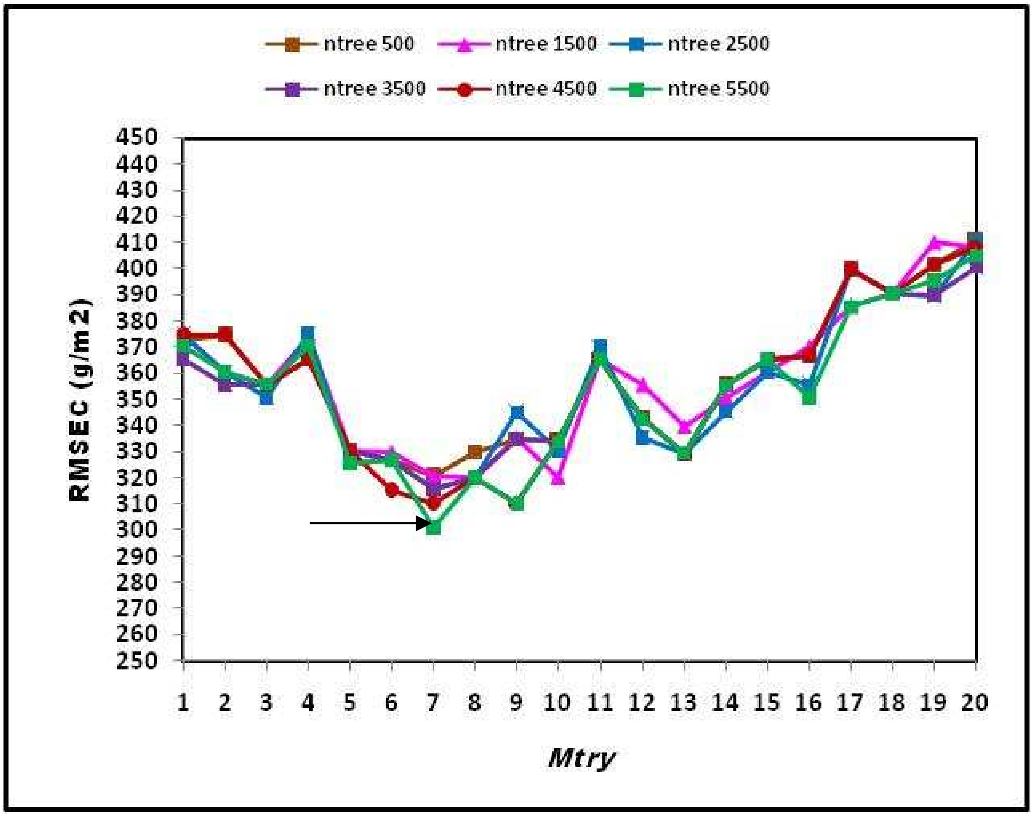

Figure 7.2. Determining the best random forest parameters (mtry and ntree) as determined by the root mean square error of prediction (RMSEP). The black arrow shows the lowest RMSEC value.

7.3.3 Determination of predictor variables

In order to simplify the modelling process, it was necessary to identify the smallest number of NDVIs that would offer the best predictive performance for AGB. The RF procedure measured the importance of the top 20 combinations from the training dataset (70 %) based on the mean decrease in the accuracy (Figure 7.3). The variables were ranked according to their importance. We subsequently used this ranking to identify the sequence in which to discard the least important variables (NDVIs) using the backward elimination function.

Figure 7.4 shows the results of the variables selection using the backward elimination function. It is interesting to note that the RMSEC generally decreased while the least important variables were discarded progressively by the backward elimination method. The best model developed using four NDVIs achieved the lowest RMSEC using the OOB sample (269 g/m^2), 10 fold cross- validation (271 g/m^2) and hold out test dataset (276 g/m^2). These four NDVIs involve a combination of wavelengths located in the NIR (853 nm, 853 nm, 847 nm, and 776 nm) and shorter wavelengths of the red-edge (741 nm, 740 nm, 741 nm, and 749 nm) respectively.

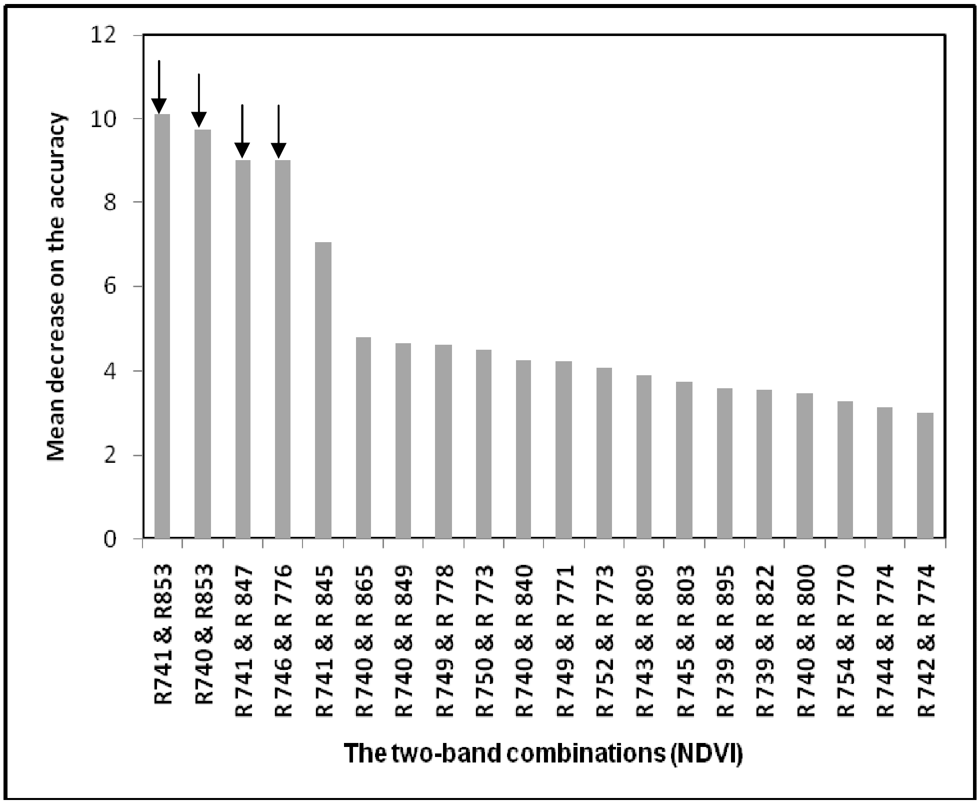

Figure 7.3. Variables importance measurement determined by OOB from the training dataset using the RF algorithm with default setting of *mtry* and 5500 *ntree*. The most important variables are shown by black arrows.

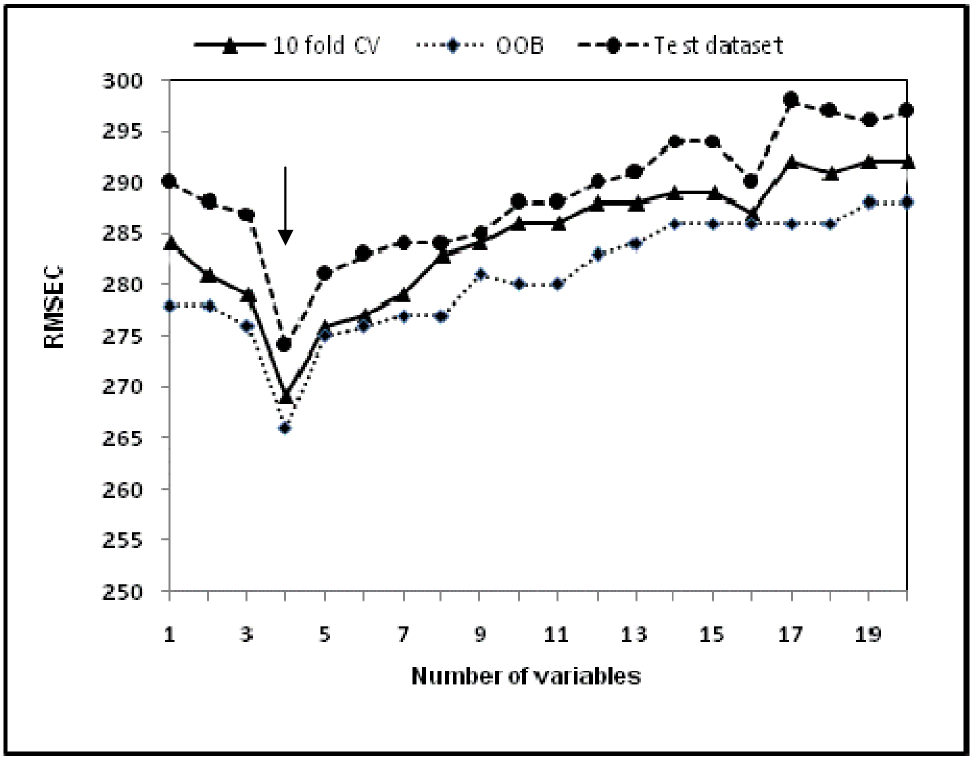

Figure 7.4. The optimal predictive variables selection using the backward elimination process. The RMSEC is calculated from the training dataset (n = 33) using OOB method, 10 fold cross-validation, and the test dataset (n = 14). The lowest RMSEC obtained is shown by the black arrow.

7.3.4 Development of the prediction model

The selected four narrow band NDVIs were used to test the performance of the RF regression in predicting the above ground biomass. Table 7.3 shows the RF prediction performance of the best selected NDVIs (n = 4) based on the coefficient of determination and root mean square error for calibration and validation. The R^2 values and root mean square error for calibration (n = 32) and test (n = 14) datasets indicate the best predictive performance of the RF model obtained when using the selected four NDVIs: NDVI (853 nm, 741nm), NDVI (853 nm, 740 nm), NDVI (847 nm, 741 nm), and NDVI (749 nm, 76 nm).

The performance of the best selected NDVIs (n = 4) was compared to those obtained by the standard NDVI calculated from a near-infrared (833 nm) and red band (680 nm) (Tucker, 1977), the best NDVI computed in this study (853 nm and 741 nm), and the top 20 NDVIs listed in Table 7.2. Where Table 7.3 and Figure 7.5 clearly show that the regression model involving the combination of the best four NDVIs yielded the highest R^2 (0.77) for the calibration and R^2 (0.73) for the prediction as well as the lowest RMSEC (266 g/m^2 = 8.2 % of the mean) and RMSEP (276 g/m^2 = 8.6 % of the mean) compared with the top 20 NDVIs which yielded a RMSEC value of 280 g/m^2 and a RMSEP value of 305 g/m^2. The lowest R^2 (0.026) and R^{2}cv (0.015) and the highest RMSEC (539 g/m^2) and RMSEP(694 g/m^2) were obtained with the standard NDVI calculated from 833 nm and 680 nm. The poor performance of the standard NDVI can be clearly noted on the almost flat scatter plot in Figure 7. 5-A for both calibration (n = 32) and independent validation (n = 14).

Table 7.3: The performance of the random forest model for prediction of papyrus biomass in the Greater St Lucia Wetland Park using different subsets of NDVIs

	Calibration (n = 33)			Independent validation (n = 14)		
	R^2 actual vs. Predicted	RMSEC g/m2	Mean %	R^2 actual vs. predicted	RMSEP g/m	Mean %
Standard NDVI(833nm and 680 nm)	0.026	539	16.7	0.015	694	21.5
Best NDVI (741 nm and 853 nm)	0.72	295	9.2	0.66	306	9.5
Selected NDVIs (n = 4)	0.77	266	8.2	0.73	276	8.6
Top 20 NDVIs	0.69	301	9.3	0.66	312	9.7

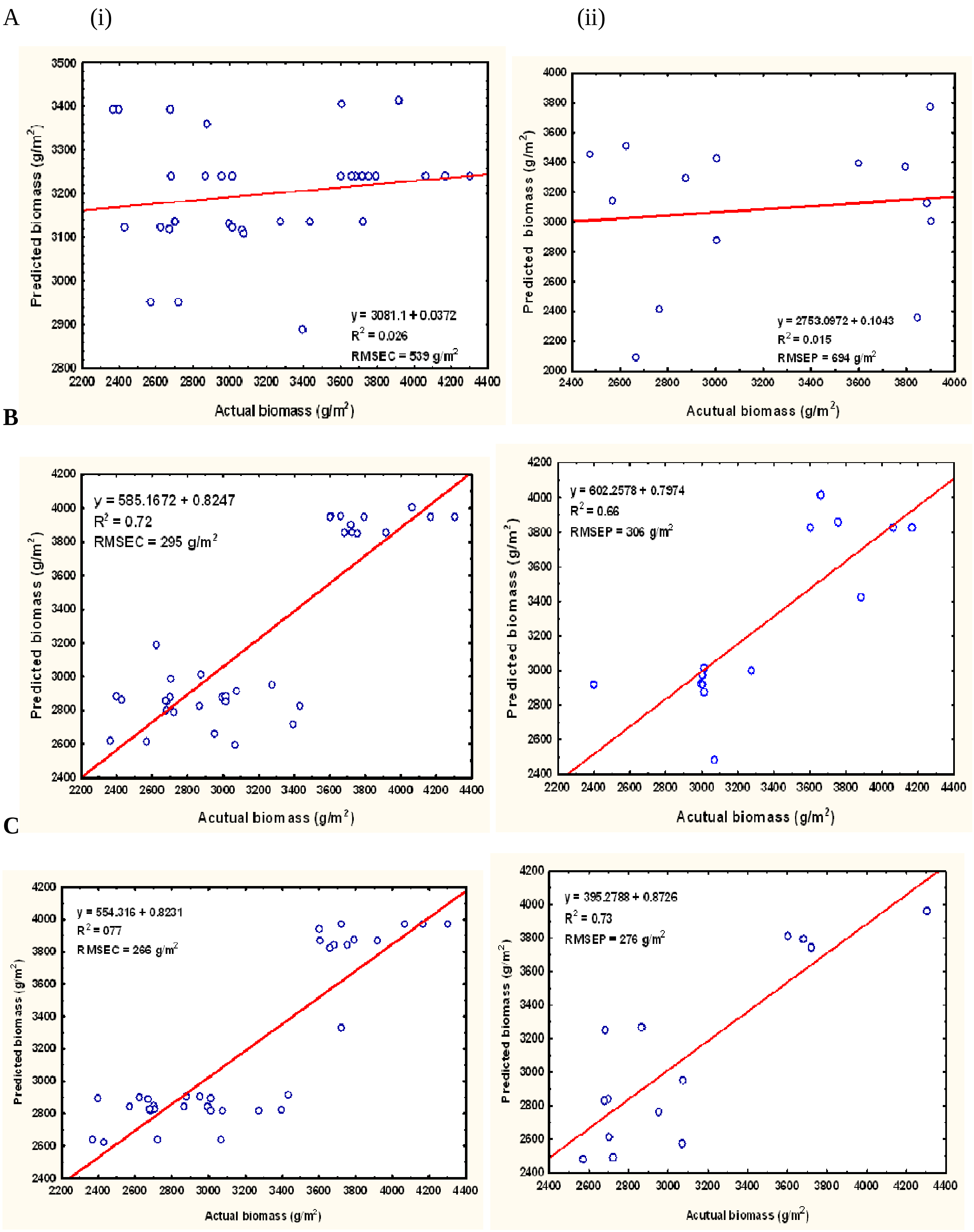

Figure 7.5. One-to-one relationships between actual and predicted papyrus biomass for (i) calibration (n = 32) and (ii) independent validation (n = 14). Random forest was developed using (a) the standard NDVI computed from a near-infrared band (833 nm) and red band (680 nm), (b)

the best narrow band NDVI developed in this study computed from 853 nm and 741 nm, and (c) the best four narrow band NDVIs computed from (853 nm and 741 nm), (853 nm and 740 nm), (847 nm and 741 nm), and (776 nm and 749 nm). For each model, R^2, RMSEC, and RMSEP are reported.

7.4 Discussion

The use of remote sensing techniques in estimating biomass from dense vegetation or high leaf area index (LAI) has been constrained by the asymptotic saturation of vegetation indices such as NDVI (Tucker, 1977; Kumar *et al.*, 2001; Mutanga and Skidmore, 2004a). This is particularly true for wetland environments, where the vegetation grows very densely (Li *et al.*, 2007). Therefore, there is an increase in NIR region reflectance due to multiple scattering effects while the absorption in the red region between 660 nm and 680 nm reaches a peak (Kumar *et al.*, 2001). This imbalance between saturation of red light absorption and high NIR reflectance causes the poor performance of the widely used remotely sensed indices such as broad band NDVI in estimating the wetland biomass because in such situations the NDVI reflects mainly canopy properties rather than the trunk properties (Tucker, 1977; Li *et al.*, 2007). The present study showed that papyrus biomass can be estimated with remarkable accuracy in areas of high dense vegetation using the RF regression algorithm and a narrow band NDVI calculated from the red-edge and NIR regions of electromagnetic spectrum.

7.4.1 Relationship between the narrow band NDVIs and biomass

The model developed in this study indicated that there is considerable information on the status of papyrus biomass contained in the red-edge and near-infrared wavelengths. The narrow band NDVI combinations calculated from these wavelengths resulted in a relatively wide variation in R^2 values (0.0 to 0.82) for estimating papyrus biomass. However, the high correlation between AGB and NDVIs obtained in this study (Table 7.2) consisted of a narrow band NDVI calculated from shorter wavelengths of the near-red-edge portion of the electromagnetic spectrum (700 nm to 750 nm), which is associated with change in chlorophyll content (Filella and Penuelas, 1994; Lichtenthaler *et al.*, 1996a), and the longer wavelengths of the red edge (750 nm to 800 nm). This result is consistent with the findings of previous studies (Mutanga and Skidmore, 2004a; Cho *et al.*, 2007; Cho and Skidmore, 2009; Fava *et al.*, 2009). Additionally, the wavelengths used to develop the best NDVIs (n = 20) (Table 7.2) in this study are within ± 10 nm of the known wavelengths that have strong relationships with biomass prediction as reported in other studies. These are 740 nm, (Cho *et al.*, 2007), 746 nm (Mutanga and Skidmore, 2004a), and 775 nm (Kawamura *et al.*, 2008).

7.4.2 Parameter optimization of the random forest model

In recent years the RF algorithm has proven to be a powerful classification method in the field of remote sensing (Gislason *et al.*, 2006; Lawrence *et al.*, 2006; Adam *et al.*, 2009). Our remarkable results from this study confirm the utility of RF as a robust, unbiased measure of error rate and an accurate regression approach for predicting biomass (Grimm *et al.*, 2008; Ismail and Mutanga, 2009).

In order to improve the prediction performance of the RF algorithm, it was first necessary to optimize the settings of the RF parameters (*ntree* and *mtry*) using the RMSEC (Breiman, 2001; Grimm *et al.*, 2008). We used all the possible values for *mtry* (The default value is one-third of the total number of variables), while the interval value of 1000 trees was used for optimizing *ntree* (the default setting of the *ntree* is 500). The results of this study revealed that the lowest RMSEC could be achieved using the default *mtry* values. This is consistent with previous studies (Liaw and Wiener, 2002; Díaz-Uriarte and de Andrés, 2006; Grimm *et al.*, 2008) which reported that the default *mtry* is often the best choice. With respect to the *ntree* optimization, the results of this study showed better predictive performance of the RF model with increasing *ntree* values. This supports the assertions made in the other studies that the highest accuracy and stability of the RF algorithm can be achieved by using a large number of trees (Díaz-Uriarte and de Andrés, 2006; Adam *et al.*, 2009). This could be explained by the fact that a forest consisting of a high number of trees (*ntree*) allows for the utilization of more variables in the dataset. It is, therefore, more stable and less prone to prediction errors caused by data perturbations (Breiman, 1996; Archer and Kimes, 2008; Zhang and Wang, 2009).

7.4.3 Variables selection

It has been noted that the use of the standard NDVI might not be able to explore the strength of the large number of hyperspectral bands because only two bands from red and NIR are used to formulate the NDVI (Hansen and Schjoerring, 2003; Schlerf *et al.*, 2005). In the present study, the results of calculating the narrow band NDVIs from all possible two-band combinations between red and NIR and then correlating it with AGB (g/m^2) improved an understanding of the relationship between the wavelength regions and biomass estimation at full canopy cover, as well as presented a possibility to explore the rich information content in the hyperspectral wavelengths (Thenkabail *et al.*, 2000; Mutanga and Skidmore, 2004a). This study demonstrates the validity and significance of NDVI in estimating AGB. However, selection of the best wavelengths is an important task for the formulation of the NDVI. Our results as shown in Figure 7.1 explored and ranked all the possible wavelength combinations, then the best combination of wavelengths (n = 20) was selected based on the strong correlation with AGB for further analysis. Besides ranking and selecting the best narrow band combinations (n = 20) that yielded the highest correlation with biomass, using the RF algorithm with backward elimination search function facilitated the selection of the fewest most important predictive variables (n = 4) for a simple modelling process and best predictive accuracy. The consistency of the three methods (OOB, 10 fold cross-validation, and the test dataset) proposed in this study to identify the optimal number of the predictive variables (n = 4) demonstrates the reliability of OOB as an internal estimate of error rate in the RF algorithm. Our finding in this regard is identical to those of other studies that tested the reliability of the OOB estimate error in the classification model (Lawrence *et al.*, 2006; Adam and Mutanga, In review) and the regression model (Ismail and Mutanga, 2009).

7.4.4 The predictive performance of the random forest model

The present study showed that papyrus biomass can be estimated at full canopy level in complex swamp wetland environments using narrow band NDVIs derived from spectrometry data and the RF regression algorithm. The higher accuracy obtained in this study demonstrated the utility of the RF algorithm as a feature selection method (Lawrence *et al.*, 2006; Adam *et al.*, 2009) and its application as a regression model (Ismail and Mutanga, 2009). The relatively high R^2 and low RMSEC and RMSEP as shown in Table 7.3 indicates that the selected NDVIs (n = 4) improved the predictive performance of the model compared to the use of the entire top 20 NDVIs. The increase in number of predictive variables could lead to a decrease in the model accuracy because the noise in the redundant data propagates through the model performance (Bajcsy and Groves, 2004). Our results in this regard indicate that the variables selection method developed in this study was able to refine the performance of the RF regression model. The poor predictive performance of the standard NDVI shown in Figure 7.5 is consistent with the finding of Cho et al. (2007), involving grass/herb in the Majella National Park in Italy, and of Mutanga and Skidmore (2004a), involving blue buffalo grass (*Cenchrus Ciliaris*) grown under controlled conditions in a greenhouse. This could be explained by the saturation problem of the standard NDVI at the high biomass or leaf area index which has been reported in several studies (Tucker, 1977; Mutanga and Skidmore, 2004a).

In summary, the RF regression model was able to provide remarkable accuracy in estimating biomass in wetland areas. The potential use of this method which was developed from fine spectral resolution (ASD) can be made operational by further work to explore the capability to estimate papyrus biomass using relatively coarser spectral data such as Hyperion or the HYMAP spectra.

7.5 Conclusions

We conclude that:

1. NDVI computed from a combination of narrow band in the shorter wavelengths of red-edge or far-red (700 nm-750 nm) and the longer wavelengths of NIR (750 nm -1000 nm) perform better in predicting biomass as compared to the standard NDVI when there is high canopy density.
2. The RF ensemble reduced the redundancy of hyperspectral data and simplified the modelling process by identifying the optimal number of narrow band NDVIs that offer the best predictive accuracy.
3. Based on our relatively high accuracies, it is worth considering the RF ensemble as a robust method for remote sensing regression type applications in the future. Our study offers the foundation for the possible upscaling of these results to coarser spectral data such as Hyperion or the HYMAP image data.

Overall, this study has revealed that it is possible to predict dense papyrus biomass at canopy level using filed spectrometry measurements. In addition, the developed model provides a better understanding of (i) those narrow band regions that are most sensitive for papyrus biomass estimation and (ii) the potential of RF ensemble as a feature selection and regression type model in remote sensing applications This permits the upscaling of the model to spaceborne or airborne sensors such as HYMAP and Hyperion.

Acknowledgements

We thank GSLWP's authority for allowing us to access the park and for providing excellent conditions for field data collection. Many thanks go to Khalid Manssour and Faiz Awad for their support in the field data collection. We thank the University of KwaZulu-Natal for providing the optimal conditions for the success of the study and assistance with the ASD. We thank Elfashir University as well as the NRF for partly funding this study. Gratitude is extended to the R Development Team for their very powerful open source package for the statistical analysis.

CHAPTER EIGHT

Remote sensing of papyrus vegetation (*Cyperus papyrus L.*) in a swamp wetland: A synthesis

8.1 Introduction

Why do we need to map and monitor papyrus (*Cyperus papyrus L.*)? Research in wetland ecology and management has revealed that *Cyperus papyrus L.* is the most important species in tropical African wetlands that plays fundamental ecological, hydrological, and economic roles in the tropical African wetlands (Grenfell *et al.*, 2009). The existence of papyrus, however, is threatened by human encroachment and intensive agricultural activities in many tropical African wetlands (Maclean *et al.*, 2006; Owino and Ryan, 2007).Therefore, detecting and monitoring the existence and quantity (biomass) of papyrus at fine spatial scales is critically important for the wetland manager and decision makers when implementing effective wetland management practices. In this regard, remote sensing is widely viewed as being a near-real-time and cost-efficient technology that has the ability to spatially proceed with large scale detecting and monitoring of the vegetation parameters.

However, detecting and mapping wetland plants such as papyrus is challenging for two reasons. Firstly, herbaceous wetland vegetation exhibits high spectral and spatial variability because of the steep environmental gradients which produce short ecotones and sharp demarcations between the vegetation units (Schmidt and Skidmore, 2003; Adam and Mutanga, 2009; Zomer *et al.*, 2009). Hence, it is often difficult to identify the boundaries between vegetation community types. Secondly, the reflectance spectra of wetland vegetation canopies are often very similar and are combined with the reflectance spectra of the underlying soil, hydrologic regime, and atmospheric vapour (Guyot, 1990; Malthus and George, 1997; Yuan and Zhang, 2006). This combination further complicates the optical classification and results in a decrease in the spectral reflectance, especially in the near-to mid-infrared regions where water absorption is relatively stronger (Fyfe, 2003; Silva *et al.*, 2008). Another problem that limits the ability of remote sensing to map papyrus quantity (biomass) is that the use of remotely sensed indices such as NDVI calculated from the broad band has been bedeviled by the saturation problem at high canopy density and after certain biomass and LAI measurement (Thenkabail *et al.*, 2000; Mutanga and Skidmore, 2004a). The challenge is, therefore, to develop techniques that can focus on mapping papyrus and predicting accurately its quantity (biomass) at canopy level.
In this thesis, the objectives were:

1. To explore the usefulness of *in situ* spectroscopic data in discriminating papyrus vegetation from its co-existing species (binary class techniques),
2. To investigate the usefulness of *in situ* spectroscopic data in discriminating among papyrus vegetation and its co-existing species (multiclass techniques),
3. To determine if machine learning algorithms (random forest) can accurately discriminate among papyrus and other co-existing species using resampled HYMAP data,
4. To examine whether vegetation indices derived from spectroscopy data can be used to enhance the separability and classification accuracy between vegetation species,
5. To test the reliability and robustness of the internal accuracy assessment of the RF algorithm as a variable selection and classification algorithm in discriminating between the species,
6. To investigate the potential of imaging spectroscopy in discriminating among papyrus and its co-existing species using airborne hyperspectral data (AISA eagle), and
7. To explore the potential of hyperspectral data in estimating biomass of papyrus at high canopy density or full canopy levels.

8.2 Spectral discrimination of papyrus under full canopy cover

In hyperspectral remote sensing of vegetation there are two different schools of thought regarding the possibility of species discrimination: the believers and the sceptics. The sceptics argue that several species may actually have a quantitatively similar spectrum which is a mixture of physical and chemical properties that can change according to various environmental factors and therefore the uniqueness of the vegetation spectra is questionable (Anderson, 1970; Price, 1994; Portigal *et al.*, 1997). Moreover, this spectral reflectance is controlled by a limited number of independent variables such as chlorophyll *a*, chlorophyll *b* and the carotenoids in the visible regions. Therefore, they argue that the reflectances of vegetation of different species are highly correlated (Price, 1992; Danson and Plummer, 1995).

On the other hand, the group of scientists who believe that spectral reflectance can be used to discriminate species has argued strongly that despite the challenges and the non-unique nature of the spectral response, the potential for discriminating different plant species based on foliar reflectance does exist because the spectral response still provides enough information to discriminate between the species (Cochrane, 2000). Furthermore, hyperspectral remote sensing enables the quantification of all the independent variables mentioned by Price (1992), such as chlorophyll content of plants (Blackburn and Pitman, 1999), biochemical variables such as nitrogen and lignin (Curran *et al.*, 1990; Mutanga, 2005), crop moisture variations (Penuelas *et al.*, 1993a), and leaf pigment concentrations (Blackburn, 1998).

In this thesis we have attempted to answer the question: can the papyrus plant be discriminated from its co-existing species in two discrimination levels; one to discriminate papyrus from each one of its co-existing species (binary class classification), and to discriminate among papyrus and its co-existing species (multi-class classification). The binary class focused on discriminating papyrus and the broader co-existing species, while the multi class focused on detailed discrimination for papyrus and its co-existing species. This allowed one to test the influence of the level of discrimination detail in the classification accuracy.

8.2.1 Spectral discrimination of papyrus from its co-excising species (*binary class*)

The evaluation of hyperspectral data (350 nm to 2500 nm) measured in the field at full canopy level shows that we can successfully discriminate papyrus from each one of its co-existing species (binary class classification) (Chapter 3). The utility of a new hierarchical method that integrates three analysis levels (ANOVA, CART, and distance analysis) indicates that there is a significant difference ($p < 0.001$) between the mean spectral reflectance for papyrus and the three co-existing species, with a large number of significant wavelengths (n= 412) located in the near-infrared and red-edge regions of the electromagnetic spectrum .The majority of the significant bands (98 %) are located in the near-infrared part (982 nm to 1297 nm) of the electromagnetic spectrum, and the remainder of the significant wavelengths are located in the red-edge part. CART analysis was able to identify the most sensitive bands for the spectral discrimination. Specifically, these bands are located in the red-edge and near-infrared region at 745 nm, 746 nm, 892 nm, 932 nm, 934 nm, 958 nm, 961 nm, 989 nm. The sensitivity analysis involving Jeffries-Matusita (JM) distance was then used to determine the best combinations of these bands for discriminating papyrus from its co-existing species. Results show that, although a single band located in 892 nm can discriminate Cyperus papyrus from *Phragmites australis* and *Thelypteris interrupta*, with JM value of 1.409 (99.64 %) and 1.408 (99.58 %) respectively, only six bands located at 745 nm, 746 nm, 892 nm, 934 nm, 958nm, and 961nm, show the potential to discriminate *Cyperus papyrus* from *Echinochloa pyramidalis* with a JM value of 1.379 (97.52)

(Table 8.1). This six band combination produces an acceptable JM separability (99.15 %) for the discrimination of papyrus from all the three co-existing species.

Table 8.1: The values of the JM distance for each individual class pair within the selected best band combinations

Best combination	CP vs PA		CP vs EP		CP vs TI	
	JM value	%	JM value	%	JM value	%
892.	1.409	99.64	1.210	85.57	1.408	99.58
892, 934.	1.412	99.86	1.263	89.32	1.410	99.72
892, 934, 898.	1.413	99.93	1.308	92.50	1.413	99.93
892,934, 958, 961,	1.414	100.00	1.329	93.99	1.414	100.00
745, 745, 892, 958, 961.	1.414	100.00	1.351	95.55	1.414	100.00
745,745, 892, 934, 958, 961.	1.414	100.00	1.379	97.52	1.414	100.00
745, 746, 892, 932, 958, 961, 989.	1.414	100.00	1.399	98.94	1.414	100.00
745, 746, 892, 932, 934, 958, 961, 989.	1.414	100.00	1.405	99.36	1.414	100.00

The results from this study provide the basis for future powerful algorithms that can be used to discriminate among papyrus and the three co-existing species (multi-class classification) at full canopy level.

8.2.2 Spectral discrimination of papyrus and its co-existing species (multi-class classification)

We assessed the potential of discriminating among papyrus and the different co-existing species (multi-class classification) using machine learning algorithms (random forest) and canopy reflectance measured under field conditions and resampled to HYMAP resolution (Chapter 4). The approach of using a wrapper (forward variable selection) and .632+ bootstrap method in tandem with the RF algorithm was able to provide small sets of non-redundant wavelengths while preserving higher classification accuracy than the full HYMAP wavelengths (n = 126)

(Table 8.2). More specifically, 10 of the HYMAP wavelengths located at 1409 nm, 710 nm, 437 nm, 464 nm, 452 nm, 1424 nm, 725 nm, 480 nm, 587 nm, and 603 nm have the greatest potential for discriminating among all classes (n = 6) involving papyrus and the different co-existing species. The RF algorithm also yielded better classification accuracy (88.44%) than the classification tree (CT) algorithm (80.47%) when the full data set (126 wavelengths) was used (Table 8.3).

Table 8.2: The confusion matrix showing the classification error obtained for discrimination amongst all possible species combinations (n = 6). *Cyperus papyrus* (CP), *Echinochloa pyramidalis* (EP), *Phragmites australis* (PA), and *Thelypteris interrupta* (TI). The confusion matrix includes the accuracy between classes (ACC), KHAT statistic, producer accuracy (PA), and user accuracy (UA).

Classes	ACC %	KHAT	PA %		UA %		Row totals	Column totals
			Presence	Absence	Presence	Absence		
CP *vs* EP	96.70	0.93	95.74	97.73	97.83	95.56	91	91
CP *vs* TI	97.89	0.96	97.83	97.96	97.83	97.96	95	95
CP *vs* PA	93.75	0.88	93.75	93.75	93.75	93.75	96	96
EP *vs* PA	96.81	0.94	97.73	96.00	95.56	97.97	94	94
EP *vs* TI	94.62	0.89	95.56	93.75	93.48	95.74	93	93
PA *vs* TI	100.00	1.00	100.00	100.00	100.00	100.00	93	93
All classes	90.50	0.87	88.24	91.49	90.00	86.00	200	200

Table 8.3: The misclassification error for both the machine learning models (RF and CT) using the .632+ bootstrap method for error estimates and accuracy assessments using the top 10 wavelengths selected by RF and a full data set (126 wavelengths).

Algorithm	Top 10 wavelengths			Full data set		
	Misclassification error %	Overall Accuracy	KHAT %	Misclassification error	Overall accuracy	KHAT %
RF	8.95	90.5	87	9.19	88.44	85
CT	12.05	84.5	80	13.75	80.47	78

Our findings in this study proved that the RF algorithm is a robust and accurate method for the combined purpose of variables selection and for the classification of hyperspectral data in an application where (i) the number of samples is limited (n < p), and where (ii) vegetation species have similar spectral characteristics affected by underlying wet soil and hydrology regime. However, more investigation is required to test the reliability and stability of the RF algorithm (Lawrence *et al.*, 2006).

8.3 Improving the spectral discrimination of papyrus vegetation

The problems of the high dimensionality of hyperspectral remote sensing, the small and high correlated absorption features present in the plants spectra, and the background effects (Price, 1992; Danson and Plummer, 1995), were addressed in this thesis (Chapter 5) by evaluating the potential of vegetation indices in discriminating papyrus and its co-existing species (Filella and Penuelas, 1994; Qi *et al.*, 1995; Green *et al.*, 1997; Haboudane *et al.*, 2002; Stimson *et al.*, 2005; Cho *et al.*, 2008; Darvishzadeh *et al.*, 2008).

We tested the utility of using narrow band vegetation indices to improve the spectral separability among papyrus and its co-existing species and the classification accuracy. The utility of widely used vegetation indices particularly, NDVIs and SRs, involving all possible two-band combinations of the 20 most important bands as determined by the RF algorithm were tested. In addition, we evaluated a number of hyperspectral indices (n = 48) that were previously demonstrated to estimate plant parameters. The key finding presented in this chapter is that spectral separability among papyrus vegetation and its co-existing species may be improved from 90.5 % overall accuracy (Chapter 4) to 96 % overall accuracy by 5 optimal vegetation indices (Table 8.4). Three of these indices were published in the literature (Plant Senescence Reflectance Index, Blue/Green Index 1, and Pigment Index 4) while the other two optimal indices were obtained from the modified NDVIs involving a combination of a narrow band in the red portion (655 nm) with two wavelengths in the red-edge position (697 nm, and 705 nm). Based on relatively high overall accuracy (96 %), the use of hyperspectral indices may be considered as a new approach for discriminating plant species.

Table 8.4: Accuracies assessment for the OOB estimates and independent test data set based on the top five vegetation indices and the full data set (n = 68). The assessment includes the kappa statistic, overall accuracy (ACC), producer accuracy (PA), and user accuracy (UA).

	Top five vegetation indices								Full data set (68 vegetation indices)							
	Out-of-bag accuracy assessment				Independent accuracy assessment				Out-of-bag accuracy assessment				Independent accuracy assessment			
Classes	ACC %	Kappa	PA %	UA %	ACC %	Kappa	PA %	UA %	ACC %	Kappa	PA %	UA %	ACC %	Kappa	PA %	UA %
CP vs EP	93.7	0.87	95.7	91.7	94.4	0.89	92.6	96.2	92.2	0.84	95.4	89.1	98	0.96	96.2	100
CP vs TI	99	0.98	100	97.8	93.3	0.86	100	86.2	98.9	0.98	100	97.6	89.8	0.80	92.6	86
CP vs PA	99	0.98	100	97.8	100	1.00	100	100	98.3	0.83	89.1	93.2	94.3	0.89	92.6	96
EP vs PA	100	1.00	100	100	100	1.00	100	100	100	1.00	100	100	100	1.00	100	100
EP vs TI	95.9	0.92	97.8	93.8	96.6	0.93	100	92.9	91.3	0.83	95.4	87.2	92.6	0.85	100	86
PA vs TI	100	1.00	100	100	100	1.00	100	100	100	1.00	100	100	94.6	0.89	90.3	100
All classes	96	0.91	97.00	89	94.5	0.91	93.6	84.3	88	0.84	85	82	85.8	0.81	83	83

8.4 Airborne hyperspectral remote sensing of papyrus vegetation

The last aspect in this thesis was to scale up the method applied as discussed in the previous Chapters (3, 4, and 5) to an airborne hyperspectral sensor to discriminate among papyrus and its co-existing species.

We tested the potential use of AISA eagle data to discriminate between papyrus and its co-existing species (Chapter 6). AISA eagle scenes were acquired in February 2009 over a section of the study area (the Dukuduku forest and Futululu forest). The images were collected with 2 m spatial resolution, 272 wavebands (393 nm – 994 nm), and 2.04 nm to 2.29 nm spectral resolution. Images were taken at an altitude of approximately 1000 m above ground during cloudless periods in the daytime. A RF ensemble was employed to reduce the redundancy in the complex hyperspectral AISA data and to classify papyrus and its co-existing species. The optimal vegetation indices selected (Chapter 5) were also tested to improve the discriminatory power of the hyperspectral data. The RF classification model consisted of 8 bands (739 nm, 737 nm, 721 nm, 734 nm, 541 nm, 543 nm, 416 nm, and 539 nm) and showed 80.83 % overall accuracy and kappa value of 0.74, while the classification model that included the optimal vegetation indices (Plant Senescence Reflectance Index, Blue/Green Index 1, and Pigment Index 4, NDVI (655, 705), and NDVI (655, 697)) was able to improve the overall accuracy up to 88.98 % and kappa value of 0.85 (Table 8.5, 8.6) and (Figure 8.1). The relatively high classification accuracy of the developed models demonstrated the potential of hyperspectral AISA data for discriminating the difference in the spectra among papyrus and its co-existing species.

Table 8.5: Testing the discriminatory performance of the RF classifier using the selected bands (n = 8) and the OOB method for estimating the error rate. The confusion matrix includes the overall accuracy, kappa statistic, user accuracy, and producer accuracy for *Cyperus papyrus* (CP), *Echinochloa pyramidalis* (EP), *Phragmites australis* (PA), and *Thelypteris interrupta* (TI)

Classes	CP	EP	PA	IT	Row total
CP	24	2	4	0	30
EP	4	22	4	0	30
PA	2	2	26	0	30
IT	2	3	0	25	30
Column total	32	29	34	25	120
Producer accuracy = 75.86 % User accuracy = 73.33 %					Overall accuracy = 80.83 % Kappa = 0.74

Table 8.6: Testing the discriminatory performance of the RF classifier using the selected vegetation indices (n = 5) and OOB method for estimating the error rate. The confusion matrix

includes the overall accuracy, kappa statistic, user accuracy, and producer accuracy for *Cyperus papyrus* (CP), *Echinochloa pyramidalis* (EP), *Phragmites australis* (PA), and *Thelypteris interrupta* (TI)

Classes	CP	EP	PA	IT	Row total
CP	24	2	4	0	30
EP	3	26	0	0	29
PA	2	0	27	0	29
IT	0	2	0	28	30
Column total	29	30	31	28	118

Producer accuracy = 86.67 %	Overall accuracy = 88.98 %
User accuracy = 89.66 %	Kappa = 0.85

8.5 Predicting papyrus biomass using narrow band vegetation indices

In order to better understand papyrus quantity (biomass) interactions with the spatial distribution, we evaluated the potential of using narrow band vegetation indices and the RF regression model in predicting biomass of *Cyperus papyrus L.* measured at high canopy density (Chapter 7). More specifically, the utility of the widely used NDVI involving all the possible two-band combinations in the red, red-edge, and NIR bands (i.e. 600 nm to 1000 nm) were investigated. These indices and spectrum region were selected because they are the most commonly used in estimating biomass and crop yield (Thenkabail *et al.*, 2000; Mutanga and Skidmore, 2004a; Cho *et al.*, 2007). The discrete 401 narrow bands allowed a computation of N*N = 160,801 narrow band NDVIs for biomass prediction. Results of this analysis are shown in R^2 for each two-band combinations in Figure 8.2. All possible two-band combinations applied in this study to compute NDVIs allowed exploring the strength of the large number of hyperspectral bands rather than focusing on the standard NDVI where only two bands from red and NIR are used to compute the index. On the other hand, the RF ensemble and backward feature elimination allowed for the reduction of redundancy of hyperspectral data and simplifying the modelling process used in this study by identifying the optimal number of narrow-band NDVIs that offer the best predictive accuracy. The RF algorithm was also used to develop biomass prediction models.

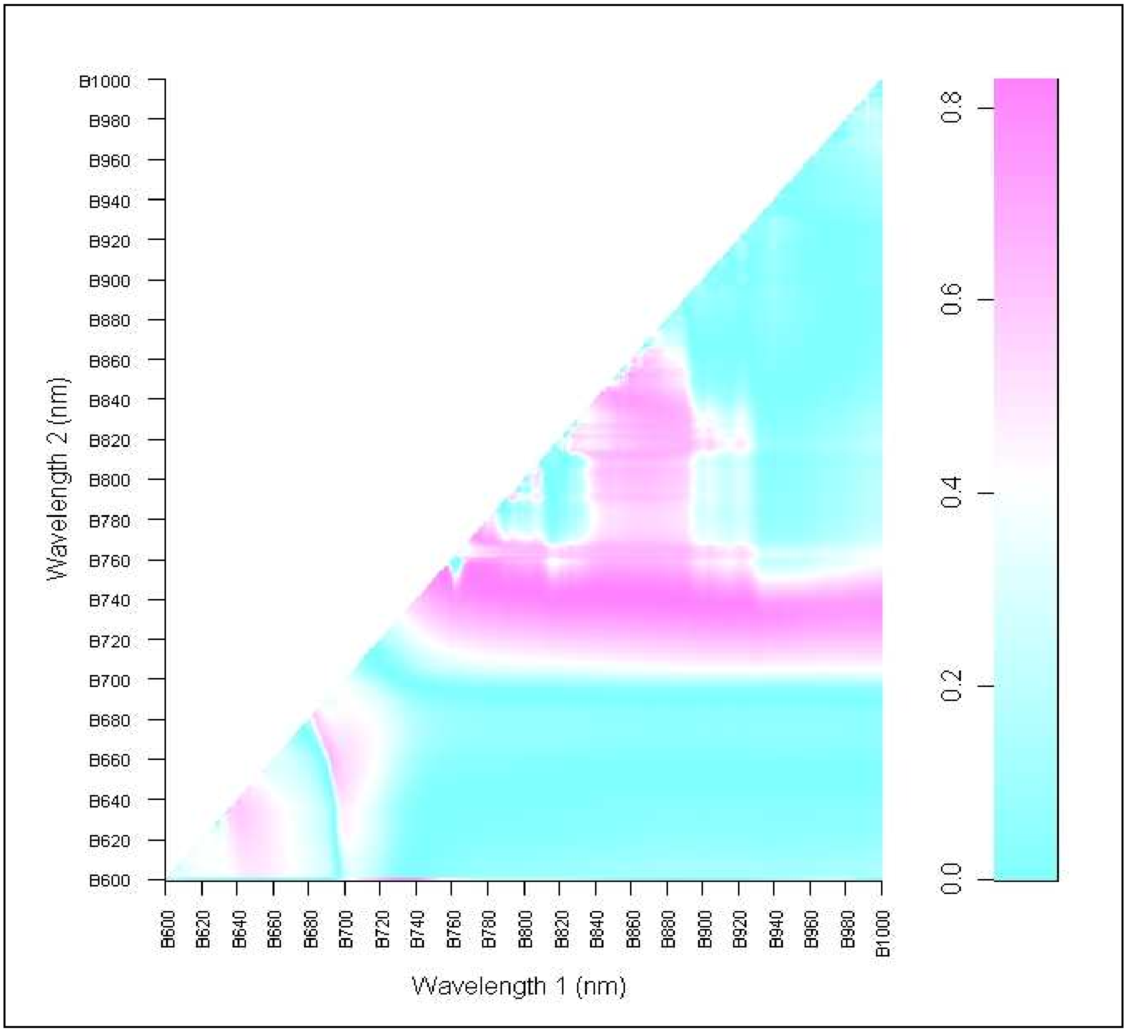

Figure 8.2. Contour plot representing the correlation coefficients (R^2) of the linear regression between above ground green biomass and NDVIs obtained from all possible two band combinations using bands located from 600 nm to 1000 nm.

Our finding in this study is that four NDVIs involving the combination of wavelengths located in the NIR (853 nm, 853 nm, 847 nm, and 776 nm) coincided with shorter wavelengths of the red-edge (741 nm, 740 nm, 741 nm, and 749 nm) respectively have the best prediction performance of papyrus biomass than the standard NDVI (833nm and 680 nm). Using these selected NDVIs (n = 4), papyrus biomass can be estimated at high canopy density (R^2 = 0.73, RMSEP = 276 g/m^2; 8.6 % of the mean) (Table 8.7).

Table 8.7: The performance of the random forest model for prediction of papyrus biomass in the Greater St Lucia Wetland Park using different subsets of NDVIs

	Calibration (n = 33)			Independent validation (n = 14)		
	R^2 actual vs. Predicted	RMSEC g/m2	Mean %	R^2 actual vs. predicted	RMSEP g/m	Mean %
Standard NDVI(833nm and 680 nm)	0.026	539	16.7	0.015	694	21.5
Best NDVI (741 nm and 853 nm)	0.72	295	9.2	0.66	306	9.5
Selected NDVIs (n = 4)	0.77	266	8.2	0.73	276	8.6
Top 20 NDVIs	0.69	301	9.3	0.66	312	9.7

In recent years, the RF has proven to be a powerful classification method in the field of remote sensing (Gislason *et al.*, 2006; Lawrence *et al.*, 2006). To the best of our knowledge, only one study by Ismail and Mutanga (2009) investigated the use of the RF algorithm in regression type applications for predicting *S. noctilio* induced water stress in *P. patula* trees using hyperspectral data. The important finding in the present studying is that the machine learning RF algorithm is deemed to be a robust, unbiased measure of error rate for feature selection. Therefore, the RF algorithm is worth considering as a robust method for remote sensing regression type applications in the future.

8.6 Evaluating the reliability and robustness of random forest algorithms for hyperspectral remote sensing classification and regression

Hyperspectral data tend to be relatively more difficult to process due to the geometrical and statistical properties associated with high dimensional data which requires sufficient training samples (Borges *et al.*, 2007; Hsu, 2007; Tsai *et al.*, 2007). Practically, in most of the hyperspectral applications, the number of training samples is limited compared to the large number of hyperspectral bands (Hsu, 2007). This is particularly true in papyrus swamps, where collecting such sufficient training and test samples is difficult due to poor accessibility. Given these problems, the challenge was to develop and test robust methods and techniques for the effective processing and classification of hyperspectral data.

In this thesis (Chapter 4 to Chapter 7) we tested the utility of the RF algorithm as a new approach for variable selection to reduce redundancy in the complex hyperspectral data set for an accurate classification and regression model. Our findings, which are consistent with other studies, indicate that the new approach outperforms other common techniques, such as classification and regression trees in that:

1. The RF algorithm can rank the importance of bands that best contribute in the classification and regression model (Lawrence *et al.*, 2006),
2. The algorithm is faster in training when compared to the ensemble methods and requires the user to specify only the number of trees to be grown (*ntree*) and the number of variables to split the nodes of individual trees (*mtry*) (Breiman, 2001; Díaz-Uriarte and de Andrés, 2006),
3. The RF algorithm can also detect outliers, which can be very useful when some of the cases may be mislabeled (Gislason *et al.*, 2006);
4. The effects of bias, variance, and instability which usually occur in other ensembles and single classification and regression trees is minimized in the RF algorithm because the multiple classification trees are constructed based on a random subset of samples derived from the training data which then vote by plurality on the correct classification (Breiman, 2001; Lawrence *et al.*, 2006),
5. The stopping rules and pruning of trees is not necessary, and the algorithm has been shown to be robust to overfitting (Pal, 2005; Granitto *et al.*, 2006; Lawrence *et al.*, 2006), and,
6. More importantly, with the RF algorithm, it is not necessary to have cross-validation or a separate accuracy assessment data set, because the OOB error rate provides an unbiased estimate of error (Lawrence *et al.*, 2006; Prinzie and Van den Poel, 2008). Our findings indicate that the internal assessment of accuracy and error rates from the RF algorithm was nearly identical to independent test data set, 10 fold cross-validation, and .632+ bootstrap for variable selection (Figure 8.1; 8.2), classification (Table 8.4), and regression models (Table 8.7) .

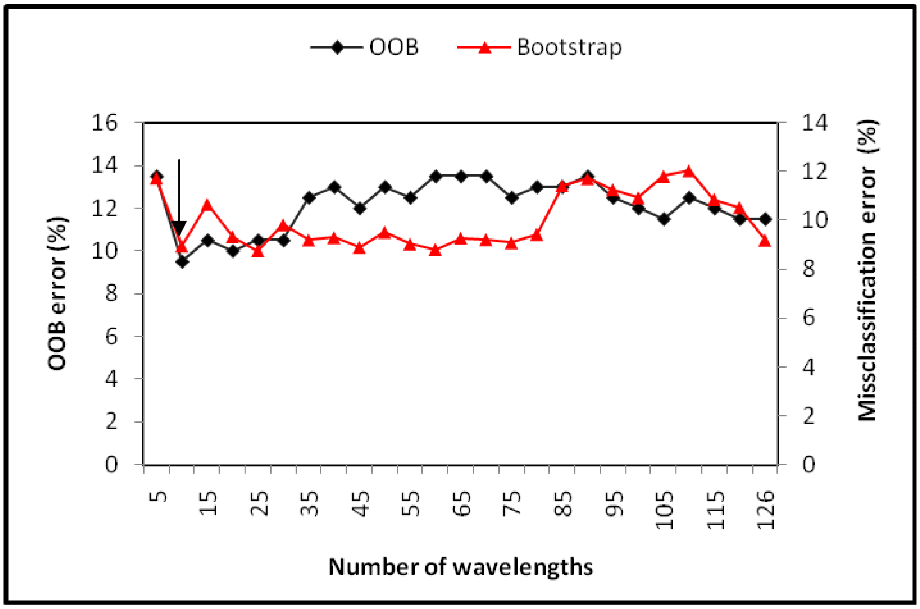

Figure 8.6. The forward variables selection method for identifying the optimal subset of wavelengths based on the OOB and .632+ bootstrap error estimates. The best subset of wavelengths with the lowest error rate is shown by the black arrow.

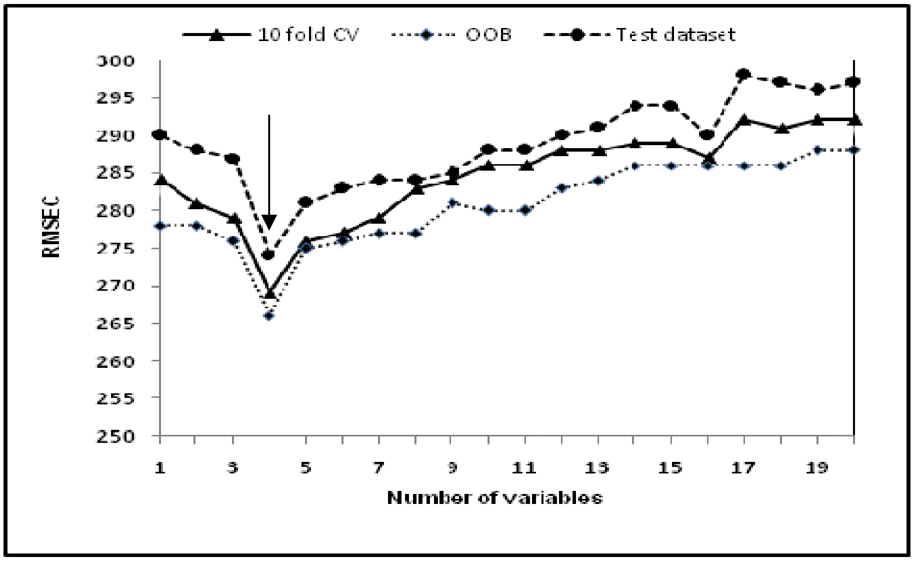

Figure 8.7. The optimal predictive variables selection using the backward elimination process. The RMSEC is calculated from the training dataset (n = 33) using OOB method, 10 fold cross validation, and the test dataset (n = 14). The lowest RMSEC obtained is shown by the black arrow.

Furthermore, an important finding in this study (Chapter 5) is that most of the overall and class accuracies based on the OOB estimates method were less than 2 % of the estimates of the independent test datasets (Table 8.4). Therefore, this combination of reliable and robust method for accuracy assessment, which obviates the need to collect a separate test dataset and of relatively high accuracies of the RF algorithm, can be considered to be desirable for hyperspectral remote sensing applications especially in complex environments such as swamp wetland areas where usually no convenient or sufficient field data are available.

8.7 conclusions

The main aim of this study was to investigate the potential of hyperspectral remote sensing techniques in discriminating spectral difference among *Cyperus papyrus L.* and three other co-existing species and in predicting biomass of *Cyperus papyrus L* in high density canopies. The findings reported in this thesis are that the information contained in hyperspectral data can accomplish these tasks. These findings contribute to the research in general and to the feasibility of applying remote sensing technologies in mapping and monitoring the distribution and the quantity (biomass) of papyrus swamps.

The main conclusions are based on the following findings from the different objectives addressed in this study:

1. Canopy reflectance measured at field level can be used to discriminate *Cyperus papyrus L.* from *P. australis, E. pyramidalis,* and *T. interrupta* (binary classification) using six wavelengths located in the red-edge and near-infrared regions of the electromagnetic spectrum. This implies that the mean spectral reflectance of *Cyperus papyrus L* is different from the other species associated with it in the same ecosystem (swamp wetlands).
2. Using the field spectrometry data resampled to HYMAP spectral resolution, the RF algorithm could also discriminate the spectral difference among *Cyperus papyrus L.* and the other co-existing species (*P. australis, E. pyramidalis,* and *T. interrupta)* (multi-class classification). This result permitted the extension of field measurements to airborne hyperspectral images for mapping papyrus and its co-existing species in swamp wetlands. The resampled data also showed the importance of the red-edge and near-infrared regions in mapping wetland plants species.

3. We have shown that hyperspectral indices can improve the spectral discrimination between papyrus and its co-existing species. Therefore, the use of narrow band indices can be considered as a new approach for discriminating wetland plants.
4. The new integrated approach developed in this study that involves the RF, as a data reduction and classification algorithm, and forward selection could discriminate among papyrus and its co-existing species with an overall accuracy of 80.98 % using airborne hyperspectral data (AISA eagle).
5. We have shown that at high canopy density, papyrus biomass could be predicted accurately using narrow band vegetation indices computed from a combination of the shorter wavelengths of red or far-red (700 nm-750 nm) and longer wavelengths of NIR (750 nm - 1000 nm), compared to the standard NDVI involving a strong chlorophyll absorption band in the red trough and a near-infrared band.
6. The machine learning RF algorithm is worth considering as a desirable technique for feature selection that can be used to reduce redundancy in the complex hyperspectral data set and that can provide powerful classification and regression applications especially in complex environments such as swamp wetland areas where usually no convenient or sufficient field data are available.

8.8 The Future

The results from this study provide an alternative method for discriminating and mapping papyrus and its co-existing species. In the future, with the operational launch of South Africa ZASat-003 satellite that will carry a hyperspectral sensor, the findings of this study will easily improve the understanding of wetland managers in developing an effective management programme for wetland ecosystems. Our findings also contribute in building the spectral libraries for different wetland plant species which will help in discriminating not only between wetland species, but also between wetland species and upland species as there has been no specific research dealing with the difference in spectral response of canopies of wetland species and upland species. Furthermore, the availability of hyperspectral sensors will allow mapping of species quality in wetland ecosystems. This includes the biochemical variables that are important in monitoring the health of papyrus swamps such as nitrogen, water content, water stress, and chlorophyll. This will help to establish a fundamental understanding of the spatial distribution of papyrus swamps functions and quality which could lead to the development of early warning systems to detect any subtle changes in the swamp systems, such as signs of stress, and could lead to the development of techniques to classify wetland area conditions (e.g. healthy or disturbed) based on their species quality and quantity.

This study focused mainly on highlighting the optimal spectral resolution for better discrimination among papyrus and other three co-existing species. In order for remote sensing methods to become operational for mapping papyrus and other species, it is critical to investigate the optimal spatial resolution and pixel size that could better map papyrus and its co-existing species in highly diverse environments. It is recommended that future research focuses on methods that consider papyrus and its co-existing species at their optimal spatial resolution (Marceau *et al.*, 1994). This will allow an increase of the information content per pixel (Atkinson, 1997).

The performance and robustness of the RF ensemble in classification models using complex hyperspectral data where the number of samples exceeds the variables (small *n* large *p*) is fully understood (Ham *et al.*, 2005; Pal, 2005; Gislason *et al.*, 2006; Lawrence *et al.*, 2006; Adam *et al.*, In press). However, to the best of our knowledge only two studies (Ismail and Mutanga, 2009; Adam, In review) examined the use of the RF algorithm in regression models using hyperspectral data. It is recommended that future studies compare the validity and reliability of the RF ensemble against other tree-based ensembles (e.g. bagging and boosting). Additionally, the RF ensemble should also be tested against other methods such as artificial neural networks which have proved to be successful in remote sensing regression model (Mutanga and Skidmore, 2004b).

REFERENCES

Abdel-Rahman, E., and Ahmed, F. (2008). The application of remote sensing techniques to sugarcane (Saccharum spp. hybrid) production: a review of the literature. *International Journal of Remote Sensing*, 29 (13), 3753-3767.

Adam, E., and Mutanga, O., Ismail, R (In review). Estimating papyrus (*Cyperus papyrus L.*) biomass using narrow band vegetation indices and the random forest regression algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Adam, E., and Mutanga, O. (2009). Spectral discrimination of papyrus vegetation (*Cyperus papyrus L.*) in swamp wetlands using field spectrometry. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64 (6), 612-620.

Adam, E., and Mutanga, O. (In review). Improving the spectral discriminating of papyrus (*Cyperus papyrus L.*) and its co-existent species at canopy level with hyperspectral indices and random forest algorithm. *International Journal of Applied Earth Observation and Geoinformation*.

Adam, E., Mutanga, O., and Rugege, D. (2010). Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. *Wetlands Ecology and Management*, 18 (3), 281-296.

Adam, E.M., Mutanga, O., Rugege, D., and Ismail, R. (2009). Field spectrometry of papyrus vegetation (*Cyperus papyrus L.*) in swamp wetlands of St Lucia, South Africa. In, *Geoscience and Remote Sensing Symposium,2009 IEEE International,IGARSS 2009* (pp. IV-260-IV-263).

Adam, E.M., Mutanga, O., Rugege, D., and Ismail, R. (In press). Discriminating the papyrus vegetation (Cyperus papyrus L.) and its co-existent species using random forest and hyperspectral data resampled to HYMAP. *International Journal of Remote Sensing*,

Aldakheel, Y., and Danson, F. (1997). Spectral reflectance of dehydrating leaves: measurements and modelling. *International Journal of Remote Sensing*, 18 (17), 3683-3690.

Anderson, J.E. (1995). *Spectral Signature of Wetland Plants (350-900)*. Alexandria: USA Army Topographic Engineering Centre.

Anderson, R. (1970). Spectral reflectance characteristics and automated data reduction techniques which identify wetland and water quality conditions in the Chesapeake Bay(Usability of multispectral, high altitude, remotely sensed data to analyze ecological and hydrological conditions in estuarine environments). In, *3 d Ann. Earth Resources Program Rev*. USA.

Andrew, M., and Ustin, S. (2006). Spectral and physiological uniqueness of perennial pepperweed (Lepidium latifolium). *Weed Science*, 54 (6), 1051-1062.

Archer, K., and Kimes, R. (2008). Empirical characterization of random forest variable importance measures. *Computational statistics & data analysis*, 52 (4), 2249-2260.

Artigas, F., and Pechmann, I.C. (2010). Balloon imagery verification of remotely sensed Phragmites australis expansion in an urban estuary of New Jersey, USA. *Landscape and Urban Planning*, 95 (3), 105-112.

Artigas, F., and Yang, J. (2005). Hyperspectral remote sensing of marsh species and plant vigour gradient in the New Jersey Meadowlands. *International Journal of Remote Sensing*, 26 (23), 5209-5220.

ASD Analytical Spectral Devices Inc. (2005). *Handheld Spectroradiometer: User's Guide, Version 4.05.*: Boulder, USA.

Asner, G.P. (1998). Biophysical and Biochemical Sources of Variability in Canopy Reflectance. *Remote Sensing of Environment*, 64 (3), 234-253.

Asrar, G. (1989). *Theory and Applications of Optical Remote Sensing*. Wiley, New York.

Atkinson, P. (1997). Selecting the spatial resolution of airborne MSS imagery for small-scale agricultural mapping. *International Journal of Remote Sensing*, 18 (9), 1903-1917.

Azza, N., Kansiime, F., Nalubega, M., and Denny, P. (2000). Differential permeability of papyrus and Miscanthidium root mats in Nakivubo swamp, Uganda. *Aquatic Botany*, 67 (3), 169-178.

Bajcsy, P., and Groves, P. (2004). Methodology for hyperspectral band selection. *Photogrammetric engineering and remote sensing*, 70 793-802.

Bajjouk, T., Populus, J., and Guillaumont, B. (1998). Quantification of subpixel cover fractions using principal component analysis and a linear programming method: application to the coastal zone of Roscoff (France). *Remote Sensing of Environment*, 64 (2), 153-165.

Bajwa, S., Bajcsy, P., Groves, P., and Tian, L. (2004). Hyperspectral image data mining for band selection in agricultural applications. *Transactions of the ASAE*, 47 (3), 895-907.

Beadle, I.C. (1974). *The Inland Waters of Tropical Africa*. Longman, New York.

Becker, B., Lusch, D., and Qi, J. (2005). Identifying optimal spectral bands from in situ measurements of Great Lakes coastal wetlands using second-derivative analysis. *Remote Sensing of Environment*, 97 (2), 238-248.

Belluco, E., Camuffo, M., Ferrari, S., Modenese, L., Silvestri, S., Marani, A., and Marani, M. (2006). Mapping salt-marsh vegetation by multispectral and hyperspectral remote sensing. *Remote Sensing of Environment*, 105 (1), 54-67.

Bemigisha, J. (2004). Remote sensing and GIS based evaluation of management options for the restoration of a papyrus swamp at lake Naivasha, Kenya. In, *The 5th AARSE conference: Geoinformation sciences in support of Africa's development* (p. 4). Nairobi

Benediktsson, J., Sveinsson, J., and Arnason, K. (1995). Classification and feature extraction of AVIRIS data. *IEEE Transactions on Geoscience and Remote Sensing*, 33 (5), 1194-1205.

Berberoglu, S., Lloyd, C., Atkinson, P., and Curran, P. (2000). The integration of spectral and textural information using neural networks for land cover mapping in the Mediterranean. *Computers & Geosciences*, 26 (4), 385-396.

Best, R.G., Wehde, M.E., and Linder, R.L. (1981). Spectral reflectance of hydrophytes. *Remote Sensing of Environment*, 11 27-35.

Blackburn, G. (1998). Spectral indices for estimating photosynthetic pigment concentrations: a test using senescent tree leaves. *International Journal of Remote Sensing*, 19 (4), 657-675.

Blackburn, G., and Pitman, J. (1999). Biophysical controls on the directional spectral reflectance properties of bracken (Pteridium aquilinum) canopies: results of a field experiment. *International Journal of Remote Sensing*, 20 (11), 2265-2282.

Boar, R.R. (2006). Responses of a fringing Cyperus papyrus L. swamp to changes in water level. *Aquatic Botany*, 84 (2), 85-92.

Borengasser, M., Hungate, W., and Watkins, R. (2007). *Hyperspectral remote sensing: principles and applications*. CRC.

Borges, J.S., Marcal, A.R.S., and Dias, J.M.B. (2007). Evaluation of feature extraction and reduction methods for hyperspectral images. In Z. Bochenek (Ed.), *New Developments and Challenges in Remote Sensing* (pp. 255-264). Poland

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24 (2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, 45 (1), 5-32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. . California, USA: Wadsworth & Brooks.

Brown, K. (2004). Increasing classification accuracy of coastal habitats using integrated airborne remote sensing. In, *EARSeL eProceedings* (pp. 34-42)

Bucci, L. (2004). *Botanical Roots of Commercial Fibbers*. Denison University.

Carpenter, G., Gopal, S., Macomber, S., Martens, S., and Woodcock, C. (1999). A neural network method for mixture estimation for vegetation mapping. *Remote Sensing of Environment*, 70 (2), 138-152.

Carter, G. (1994). Ratios of leaf reflectances in narrow wavebands as indicators of plant stress. *International Journal of Remote Sensing*, 15 (3), 697-703.

Ceccato, P., Flasse, S., Tarantola, S., Jacquemoud, S., and Grégoire, J. (2001). Detecting vegetation leaf water content using reflectance in the optical domain. *Remote Sensing of Environment*, 77 (1), 22-33.

Chan, J.C.-W., and Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112 (6), 2999-3011.

Chappelle, E., Kim, M., and McMurtrey III, J. (1992). Ratio analysis of reflectance spectra (RARS): An algorithm for the remote estimation of the concentrations of chlorophyll a, chlorophyll b, and carotenoids in soybean leaves. *Remote Sensing of Environment*, 39 (3), 239-247.

Cho, M., and Skidmore, A. (2009). Hyperspectral predictors for monitoring biomass production in Mediterranean mountain grasslands: Majella National Park, Italy. *International Journal of Remote Sensing*, 30 (2), 499-515.

Cho, M., Skidmore, A., Corsi, F., van Wieren, S., and Sobhan, I. (2007). Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. *International Journal of Applied Earth Observation and Geoinformation*, 9 (4), 414-424.

Cho, M.A., Sobhan, I., Skidmore, A.K., and de Leeuw, J. (2008). Discriminating species using hyperspectral indices at leaf and canopy scales. In, *the XXI congress:Silk road for information from imagery:* (pp. 369-376). Beijing, China: the International Society for Photogrammetry and Remote Sensing.

Clark, R. (1999). Spectroscopy of rocks and minerals, and principles of spectroscopy. *Manual of remote sensing*, 3 3–58.

Clevers, J. (1999). The use of imaging spectrometry for agricultural applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54 (5-6), 299-304.

Cochrane, M. (2000). Using vegetation reflectance variability for species level classification of hyperspectral data. *International Journal of Remote Sensing*, 21 (10), 2075-2087.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20 (1), 37.

Congalton, R., and Green, K. (2008). *Assessing the accuracy of remotely sensed data: principles and practices*. Lewis Publishers .

Cooksey, D., and Sheley, R. (1997). Noxious weed survey and mapping system. *Rangelands*, 19 (6), 20-23.

Curran, P., Dungan, J., and Gholz, H. (1990). Exploring the relationship between reflectance red edge and chlorophyll content in slash pine. *Tree Physiology*, 7 (1-2-3-4), 33.

Curran, P., Dungan, J., Macler, B., and Plummer, S. (1991). The effect of a red leaf pigment on the relationship between red edge and chlorophyll concentration. *Remote Sensing of Environment*, 35 (1), 69-76.

Cutler, D., Edwards J, T., Beard, K., Cutler, A., Hess, K., Gibson, J., and Lawler, J. (2007). Random forests for classification in ecology. *Ecology*, 88 (11), 2783-2792.

Dahlberg, A. (2005). Local resource use, nature conservation and tourism in Mkuze wetlands, South Africa: A complex weave of dependence and conflict. *Geografisk Tidsskrift, Danish Journal of Geography*, 105 (1), 1-13.

Danson, F., and Plummer, S. (1995). Red-edge response to forest leaf area index. *International Journal of Remote Sensing*, 16 (1), 183-188.

Darvishzadeh, R., Skidmore, A., Atzberger, C., and van Wieren, S. (2008). Estimation of vegetation LAI from hyperspectral reflectance data: Effects of soil type and plant architecture. *International Journal of Applied Earth Observation and Geoinformation*, 10 (3), 358-373.

Datt, B. (1999). Remote sensing of water content in Eucalyptus leaves. *Australian Journal of Botany*, 47 (6), 909-923.

Daughtry, C., and Walthall, C. (1998). Spectral discrimination of Cannabis sativa L. leaves and canopies. *Remote Sensing of Environment*, 64 (2), 192-201.

Daughtry, C., Walthall, C., Kim, M., De Colstoun, E., and McMurtreyIII, J. (2000). Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. *Remote Sensing of Environment*, 74 (2), 229-239.

Davi, H., Soudani, K., Deckx, T., Dufrene, E., Dantec, V., and François, C. (2006). Estimation of forest leaf area index from SPOT imagery using NDVI distribution over forest stands. *International Journal of Remote Sensing*, 27 (5), 885-902.

Davidson, A., Wang, S., and Wilmshurst, J. (2006). Remote sensing of grassland-shrubland vegetation water content in the shortwave domain. *International Journal of Applied Earth Observation and Geoinformation*, 8 (4), 225-236.

De'ath, G., and Fabricius, K. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81 (11), 3178-3192.

Demir, B., and Ertürk, S. (2008). Phase correlation based redundancy removal in feature weighting band selection for hyperspectral images. *International Journal of Remote Sensing*, 29 (6), 1801-1807.

Dennison, W., Orth, R., Moore, K., Stevenson, J., Carter, V., Kollar, S., Bergstrom, P., and Batiuk, R. (1993). Assessing water quality with submersed aquatic vegetation. *BioScience*, 86-94.

Denny, P. (1997). Implementation of constructed wetlands in developing countries. *Water Science and Technology*, 35 (5), 27-34.

Díaz-Uriarte, R., and de Andrés, A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7 (1), 3.

Domaç, A., and Süzen, M. (2006). Integration of environmental variables with satellite images in regional scale vegetation classification. *International Journal of Remote Sensing*, 27 (7), 1329-1350.

Elvidge, C., and Chen, Z. (1995). Comparison of broad-band and narrow-band red and near-infrared vegetation indices. *Remote Sensing of Environment*, 54 (1), 38-48.

ENVI (2006). *Environment for Visualising Images*. USA: ITT industries, Inc.

ERDAS (2005). *ERDAS Field Guide: Leica Ecosystems Geospatial Imaging* LLC.

Fava, F., Colombo, R., Bocchi, S., Meroni, M., Sitzia, M., Fois, N., and Zucca, C. (2009). Identification of hyperspectral vegetation indices for Mediterranean pasture characterization. *International Journal of Applied Earth Observation and Geoinformation*, 11 (4), 233-243.

Filella, I., and Penuelas, J. (1994). The red edge position and shape as indicators of plant chlorophyll content, biomass and hydric status. *International Journal of Remote Sensing*, 15 (7), 1459-1470.

Filippi, A., and Jensen, J. (2006). Fuzzy learning vector quantization for hyperspectral coastal vegetation classification. *Remote Sensing of Environment*, 100 (4), 512-530.

Fuentes, D., Gamon, J., Qiu, H., Sims, D., and Roberts, D. (2001). Mapping Canadian boreal forest vegetation using pigment and water absorption features derived from the AVIRIS sensor. *Journal of Geophysical Research. D. Atmospheres*, 106 33.

Fyfe, S. (2003). Spatial and temporal variation in spectral reflectance: Are seagrass species spectrally distinct? *Limnology and Oceanography*, 48 (1), 464-479.

Gao, B. (1996). NDWI--A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58 (3), 257-266.

Gao, X., Huete, A., Ni, W., and Miura, T. (2000). Optical-biophysical relationships of vegetation spectra without background contamination. *Remote Sensing of Environment*, 74 (3), 609-620.

Garcia, M., and Ustin, S. (2001). Detection of interannual vegetation responses to climatic variability using AVIRIS data in a coastal savanna in California. *IEEE Transactions on Geoscience and Remote Sensing*, 39 (7),

Gaudet, J.J. (1980). Papyrus and ecology of Lake Naivasha, National Geographic Society,Research Reports In (pp. 267–272)

Gislason, P.O., Benediktsson, J.A., and Sveinsson, J.R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters*, 27 (4), 294-300.

Gitelson, A., and Merzlyak, M. (1994). Spectral reflectance changes associated with autumn senescence of Aesculus hippocastanum L. and Acer platanoides L. leaves. Spectral features and relation to chlorophyll estimation. *Journal of Plant Physiology*, 143 (3), 286-292.

Gitelson, A., and Merzlyak, M. (1996). Signature analysis of leaf reflectance spectra: algorithm development for remote sensing of chlorophyll. *Journal of Plant Physiology*, 148 (3), 494-500.

Gitelson, A., Merzlyak, M., and Chivkunova, O. (2001). Optical Properties and Nondestructive Estimation of Anthocyanin Content in Plant Leaves¶. *Photochemistry and Photobiology*, 74 (1), 38-45.

Gitelson, A., Zur, Y., Chivkunova, O., and Merzlyak, M. (2002). Assessing Carotenoid Content in Plant Leaves with Reflectance Spectroscopy. *Photochemistry and Photobiology*, 75 (3), 272-281.

Goetz, A. (1991). Imaging spectrometry for studying earth, air, fire and water. *EARSeL Advances in Remote Sensing*, 1 (1), 3-15.

Gong, P., Pu, R., Biging, G., and Larrieu, M. (2003). Estimation of forest leaf area index using vegetation indices derived from Hyperion hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 41 (6), 1355-1362.

Gong, P., Pu, R., and Miller, J. (1995). Coniferous forest leaf area index estimation along the Oregon transect using compact airborne spectrographic imager data. *Photogrammetric engineering and remote sensing*, 61 (9), 1107-1117.

Govender, M., Chetty, K., and Bulcock, H. (2009). A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water SA*, 33 (2),

Granitto, P., Furlanello, C., Biasioli, F., and Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83 (2), 83-90.

Green, E., Mumby, P., Edwards, A., Clark, C., and Ellis, A. (1997). Estimating leaf area index of mangroves from satellite data. *Aquatic Botany*, 58 (1), 11-19.

Green, R., Eastwood, M., Sarture, C., Chrien, T., Aronsson, M., Chippendale, B., Faust, J., Pavri, B., Chovit, C., and Solis, M. (1998). Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sensing of Environment*, 65 (3), 227-248.

Grenfell, S., Ellery, W., and Grenfell, M. (2009). Geomorphology and dynamics of the Mfolozi River floodplain, KwaZulu-Natal, South Africa. *Geomorphology*, 107 (3-4), 226-240.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island--Digital soil mapping using Random Forests analysis. *Geoderma*, 146 (1-2), 102-113.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3 1157-1182.

Guyot, G. (1990). Optical properties of vegetation canopies. In M.D. Steven (Ed.), *Applications of remote sensing in agriculture* (pp. 19-44). London,: Butterworths.

Haboudane, D., Miller, J., Tremblay, N., Zarco-Tejada, P., and Dextraze, L. (2002). Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sensing of Environment*, 81 (2-3), 416-426.

Ham, J., Chen, Y., Crawford, M., and Ghosh, J. (2005). Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43 (3), 492 - 501.

Hamza, M., and Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, 75 (8), 629-643.

Han, P., Zhang, X., Norton, R.S., and Feng, Z.P. (2007). Reducing overfitting in predicting intrinsically unstructured proteins. In, *the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'2007)* (pp. 515--522.). Nanjing, China

Hansen, P., and Schjoerring, J. (2003). Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote Sensing of Environment*, 86 (4), 542-553.

Harb, R., Yan, X., Radwan, E., and Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests. *Accident Analysis & Prevention*, 41 (1), 98-107.

Hardisky, M., Gross, M., and Klemas, V. (1986). Remote sensing of coastal wetlands. *BioScience*, 36 (7), 453-460.

Harper, D. (1992). The ecological relationships of aquatic plants at Lake Naivasha, Kenya. *Hydrobiologia*, 232 (1), 65-71.

Harvey, K., and Hill, G. (2001). Vegetation mapping of a tropical freshwater swamp in the Northern Territory, Australia: a comparison of aerial photography, Landsat TM and SPOT satellite imagery. *International Journal of Remote Sensing*, 22 (15), 2911-2925.

Hawkins, D., Basak, S., and Mills, D. (2003). Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci*, 43 (2), 579-586.

He, C., Zhang, Q., Li, Y., Li, X., and Shi, P. (2005). Zoning grassland protection area using remote sensing and cellular automata modeling--A case study in Xilingol steppe grassland in northern China. *Journal of arid environments*, 63 (4), 814-826.

Hestir, E.L., Khanna, S., Andrew, M.E., Santos, M.J., Viers, J.H., Greenberg, J.A., Rajapakse, S.S., and Ustin, S.L. (2008). Identification of invasive vegetation using hyperspectral remote sensing in the California Delta ecosystem. *Remote Sensing of Environment*, 112 (11), 4034-4047.

Howland, W. (1980). Multispectral aerial photography for wetland vegetation mapping. *Photogrammetric engineering and remote sensing*, 46 87-99.

Hsu, P. (2007). Feature extraction of hyperspectral images using wavelet and matching pursuit. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62 (2), 78-92.

Huete, A. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, 25 (3), 295-309.

Huete, A.R., and Jackson, R.D. (1988). Soil and atmosphere influences on the spectra of partial canopies. *Remote Sensing of Environment*, 25 (1), 89-105.

Hunt, G. (1977). Spectral signatures of particulate minerals in the visible and near infrared. *Geophysics*, 42 (3), 501-513.

Hunt, J., E, and Rock, B. (1989). Detection of changes in leaf water content using near-and middle-infrared reflectances. *Remote Sensing of Environment*, 30 (1), 43-54.

Ismail, R. (2009). Remote sensing of forest health: The detection and mapping of Pinus patula trees infested by Sirex noctilio. In, *School of Environmental Sciences*. Pietermaritzburg, South Africa: University of KwaZulu-Natal,.

Ismail, R., and Mutanga, O. (2009). A comparison of regression tree ensembles: Predicting Sirex noctilio induced water stress in Pinus patula forests of KwaZulu-Natal, South Africa. *International Journal of Applied Earth Observation and Geoinformation*, (In Press).

Ismail, R., Mutanga, O., and Ahmed, F. (2007). Discriminating Sirex noctilio attack in pine forest plantations in South Africa using high spectral resolution data. In M. Kalacska, andA. Sanchez-Azofeifa (Eds.), *Hyperspectral Remote Sensing of Tropical and Sub-Tropical Forests* (p. 350). Rutledge, USA: Taylor and Francis: CRC Press.

Jiang, X., Tang, L., and Wang, C. (2004). Spectral characteristics and feature selection of hyperspectral remote sensing data. *International Journal of Remote Sensing*, 25 (1), 51-59.

Johnston, R., and Barson, M. (1993). Remote sensing of Australian wetlands: an evaluation of Landsat TM data for inventory and classification. *Australian journal of marine and freshwater research. Melbourne*, 44 (2), 235-252.

Jones, M. (1983a). Papyrus: a new fuel for the Third World. *Name: New Sci*,

Jones, M., and Muthuri, F. (1985). The canopy structure and microclimate of papyrus (Cyperus papyrus) swamps. *The Journal of Ecology*, 73 (2), 481-491.

Jones, M., and Muthuri, F. (1997). Standing biomass and carbon distribution in a papyrus (Cyperus papyrus L.) swamp on Lake Naivasha, Kenya. *Journal of Tropical Ecology*, 13 (03), 347-356.

Jones, M.B. (1983b). Papyrus: a new fuel for the third world *New Scientist*, 99 418-421

Junk, W. (2003). Long-term environmental trends and the future of tropical wetlands. *Environmental Conservation*, 29 (04), 414-435.

Kamaruzaman, J., and Kasawani, I. (2007). Imaging Spectrometry on Mangrove Species Identification and Mapping in Malaysia. *WSEAS Trans Biol Biomed*, 8 118-126.

Kansiime, F., Oryem-Origa, H., and Rukwago, S. (2005). Comparative assessment of the value of papyrus and cocoyams for the restoration of the Nakivubo wetland in Kampala, Uganda. *Physics and Chemistry of the Earth, Parts A/B/C*, 30 (11-16), 698-705.

Kavzoglu, T., and Mather, P. (2002). The role of feature selection in artificial neural network applications. *International Journal of Remote Sensing*, 23 (15), 2919-2937.

Kawamura, K., Watanabe, N., Sakanoue, S., and Inoue, Y. (2008). Estimating forage biomass and quality in a mixed sown pasture based on partial least squares regression with waveband selection. *Grassland Science*, 54 (3), 131-145.

Kent, M., and Coker, P. (1994). *Vegetation Description and Analysis: A practical Approach*. London: John Wiley and Sons.

Kim, D., Lee, S., and Park, J. (2006). Building lightweight intrusion detection system based on random forest. *Advances in Neural Networks-ISNN 2006*, 224-230.

Klemas, V. (2001). Remote sensing of landscape-level coastal environmental indicators. *Environmental Management*, 27 (1), 47-57.

Knipling, E. (1970). Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation. *Remote Sensing of Environment*, 1 (3), 155-159.

Kohavi, R., and John, G. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97 (1-2), 273-324.

Kokaly, R., and Clark, R. (1999a). Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. *Remote Sensing of Environment*, 67 (3), 267-287.

Kokaly, R., Despain, D., Clark, R., and Livo, K. (2003). Mapping vegetation in Yellowstone National Park using spectral feature analysis of AVIRIS data. *Remote Sensing of Environment*, 84 (3), 437-456.

Kokaly, R.F., and Clark, R.N. (1999b). Spectroscopic Determination of Leaf Biochemistry Using Band-Depth Analysis of Absorption Features and Stepwise Multiple Linear Regression. *Remote Sensing of Environment*, 67 (3), 267-287.

Kovacs, J., Flores-Verdugo, F., Wang, J., and Aspden, L. (2004). Estimating leaf area index of a degraded mangrove forest using high spatial resolution satellite data. *Aquatic Botany*, 80 (1), 13-22.

Kovacs, J., Wang, J., and Flores-Verdugo, F. (2005). Mapping mangrove leaf area index at the species level using IKONOS and LAI-2000 sensors for the Agua Brava Lagoon, Mexican Pacific. *Estuarine, Coastal and Shelf Science*, 62 (1-2), 377-384.

Kumar, L., Schmidt, K.S., Dury, S., and Skidmore, A.K. (2001). Imaging spectrometry and vegetation science. In F. van der Meer, de Jong, S.M. (Ed.), *Imaging spectrometry* (pp. 111–155.). The Netherlands: Kluwer Academic, Dordrecht.

Kyambadde, J., Kansiime, F., Gumaelius, L., and Dalhammar, G. (2004). A comparative study of Cyperus papyrus and Miscanthidium violaceum-based constructed wetlands for wastewater treatment in a tropical climate. *Water research*, 38 (2), 475-485.

Lawrence, R.L., Wood, S.D., and Sheley, R.L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment*, 100 (3), 356-362.

Lee, K., and Lunetta, R. (1995). Wetland detection methods. In J.G. Lyon, McCarthy, J. (Ed.), *Wetland and environmental applications of GIS. .* New York

Lewis Publishers.

Lehmann, A., and Lachavanne, J. (1997). Geographic information systems and remote sensing in aquatic botany. *Aquatic Botany*, 58 (3), 195-207.

Li, L., Ustin, S., and Lay, M. (2005). Application of multiple endmember spectral mixture analysis (MESMA) to AVIRIS imagery for coastal salt marsh mapping: a case study in China Camp, CA, USA. *International Journal of Remote Sensing*, 26 (23), 5193-5207.

Li, X., Yeh, A., Wang, S., Liu, K., Liu, X., Qian, J., and Chen, X. (2007). Regression and analytical models for estimating mangrove wetland biomass in South China using Radarsat images. *International Journal of Remote Sensing*, 28 (24), 5567-5582.

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. *R news*, 2 (3), 18–22.

Lichtenthaler, H., Gitelson, A., and Lang, M. (1996a). Non-destructive determination of chlorophyll content of leaves of a green and an aurea mutant of tobacco by reflectance measurements. *Journal of Plant Physiology*, 148 (3), 483-493.

Lichtenthaler, H., Lang, M., Sowinska, M., Heisel, F., and Miehe, J. (1996b). Detection of vegetation stress via a new high resolution fluorescence imaging system. *Journal of Plant Physiology*, 148 (5), 599-612.

Lillesand, T.M., and Kiefer, R.W. (2001). *Remote Sensing and Image Interpretation*. (4th ed.). New York, USA.: John Wiley and Sons, Inc.

Liu, L., Wang, J., Huang, W., Zhao, C., Zhang, B., and Tong, Q. (2004). Estimating winter wheat plant water content using red edge parameters. *International Journal of Remote Sensing*, 25 (17), 3331-3342.

Lu, D. (2006). The potential and challenge of remote sensing-based biomass estimation. *International Journal of Remote Sensing*, 27 (7), 1297-1328.

Lu, D., and Batistella, M. (2005). Exploring TM image texture and its relationships with biomass estimation in Rondônia, Brazilian Amazon. *Acta Amazonica*, 35 249-257.

Lu, D., Mausel, P., Brondízio, E., and Moran, E. (2004). Relationships between forest stand parameters and Landsat TM spectral responses in the Brazilian Amazon Basin. *Forest Ecology and Management*, 198 (1-3), 149-167.

Ma, B., Dwyer, L., Costa, C., Cober, E., and Morrison, M. (2001). Early prediction of soybean yield from canopy reflectance measurements. *Agronomy Journal*, 93 (6), 1227.

Maclean, I., Hassall, M., Boar, R., and Lake, I. (2006). Effects of disturbance and habitat loss on papyrus-dwelling passerines. *Biological Conservation*, 131 (3), 349-358.

Mafabi, P. (2000). The role of wetland policies in the conservation of waterbirds: the case of Uganda. *Ostrich*, 71 (1&2),

Malthus, T., and George, D. (1997). Airborne remote sensing of macrophytes in Cefni Reservoir, Anglesey, UK. *Aquatic Botany*, 58 (3-4), 317-332.

Marceau, D., Howarth, P., and Gratton, D. (1994). Remote sensing and the measurement of geographical entities in a forested environment. 1. The scale and spatial aggregation problem. *Remote Sensing of Environment*, 49 (2), 93-104.

May, A., Pinder, J., and Kroh, G. (1997). A comparison of Landsat Thematic Mapper and SPOT multi-spectral imagery for the classification of shrub and meadow vegetation in northern California, USA. *International Journal of Remote Sensing*, 18 (18), 3719-3728.

McCarthy, J., Gumbricht, T., and McCarthy, T. (2005). Ecoregion classification in the Okavango Delta, Botswana from multitemporal remote sensing. *International Journal of Remote Sensing*, 26 (19), 4339-4357.

Merzlyak, M., Gitelson, A., Chivkunova, O., and Rakitin, V. (1999). Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiologia plantarum*, 106 (1), 135-141.

Milton, E., Schaepman, M., Anderson, K., Kneubühler, M., and Fox, N. (2009). Progress in field spectroscopy. *Remote Sensing of Environment*, 113 S92-S109.

Mironga, J. (2004). Geographic information systems(GIS) and remote sensing in the management of shallow tropical lakes. *Applied Ecology and Environmental Research*, 2 (1), 83-103.

Mnaya, B., Asaeda, T., Kiwango, Y., and Ayubu, E. (2007). Primary production in papyrus (Cyperus papyrus L.) of Rubondo Island, Lake Victoria, Tanzania. *Wetlands Ecology and Management*, 15 (4), 269-275.

Moreau, S., Bosseno, R., Gu, X., and Baret, F. (2003). Assessing the biomass dynamics of Andean bofedal and totora high-protein wetland grasses from NOAA/AVHRR. *Remote Sensing of Environment*, 85 (4), 516-529.

Mutanga, O. (2004). Hyperspectral Remote Sensing of Tropical Grass Quality and Quantity. In (p. 195). Wageningen, The Netherlands: Wageningen University.

Mutanga, O. (2005). Discriminating tropical grass canopies grown under different nitrogen treatments using spectra resampled to HYMAP. *International Journal of Geoinformatics*, 1 (2), 21-32.

Mutanga, O., and Kumar, L. (2007). Estimating and mapping grass phosphorus concentration in an African savanna using hyperspectral image data. *International Journal of Remote Sensing*, 28 (21), 4897-4911.

Mutanga, O., and Skidmore, A. (2004a). Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing*, 25 (19), 3999-4014.

Mutanga, O., and Skidmore, A. (2004b). Integrating imaging spectroscopy and neural networks to map grass quality in the Kruger National Park, South Africa. *Remote Sensing of Environment*, 90 (1), 104-115.

Mutanga, O., and Skidmore, A. (2007). Red edge shift and biochemical content in grass canopies. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62 (1), 34-42.

Mutanga, O., Skidmore, A., and Prins, H. (2004). Predicting in situ pasture quality in the Kruger National Park, South Africa, using continuum-removed absorption features. *Remote Sensing of Environment*, 89 (3), 393-408.

Mutanga, O., Skidmore, A., and van Wieren, S. (2003). Discriminating tropical grass (Cenchrus ciliaris) canopies grown under different nitrogen treatments using spectroradiometry* 1. *ISPRS Journal of Photogrammetry and Remote Sensing*, 57 (4), 263-272.

Muthuri, F., and Kinyamario, J. (1989). Nutritive value of papyrus (Cyperus papyrus, Cyperaceae), a tropical emergent macrophyte. *Economic Botany*, 43 (1), 23-30.

Mwaura, F., and Widdowson, D. (1992). Nitrogenase activity in the papyrus swamps of Lake Naivasha, Kenya. *Hydrobiologia*, 232 (1), 23-30.

Nagendra, H. (2001). Using remote sensing to assess biodiversity. *International Journal of Remote Sensing,* 22 (12), 2377-2400.

Nagler, P., Glenn, E., and Huete, A. (2001). Assessment of spectral vegetation indices for riparian vegetation in the Colorado River delta, Mexico. *Journal of arid environments*, 49 (1), 91-110.

Numata, I., Roberts, D., Chadwick, O., Schimel, J., Galvão, L., and Soares, J. (2008). Evaluation of hyperspectral data for pasture estimate in the Brazilian Amazon using field and imaging spectrometers. *Remote Sensing of Environment*, 112 (4), 1569-1583.

Owino, A., and Ryan, P. (2007). Recent papyrus swamp habitat loss and conservation implications in western Kenya. *Wetlands Ecology and Management*, 15 (1), 1-12.

Ozesmi, S., and Bauer, M. (2002). Satellite remote sensing of wetlands. *Wetlands Ecology and Management,* 10 (5), 381-402.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26 (1), 217-222.

Pal, M. (2006). Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data. *International Journal of Remote Sensing*, 27 (14), 2877-2894.

Pal, M., and Mather, P. (2004). Assessment of the effectiveness of support vector machines for hyperspectral data. *Future Generation Computer Systems,* 20 (7), 1215-1225.

Palmer, D., O'Boyle, N., Glen, R., and Mitchell, J. (2007). Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling*, 47 (1), 150-158.

Peña-Barragán, J., Lopez-Granados, F., Jurado-Expósito, M., and Garcia-Torres, L. (2006). Spectral discrimination of Ridolfia segetum and sunflower as affected by phenological stage. *Weed Research*, 46 (1), 10-21.

Pengra, B.W., Johnston, C.A., and Loveland, T.R. (2007). Mapping an invasive plant, Phragmites australis, in coastal wetlands using the EO-1 Hyperion hyperspectral sensor. *Remote Sensing of Environment*, 108 (1), 74-81.

Penuelas, J., Baret, F., and Filella, I. (1995). Semi-empirical indices to assess carotenoids/chlorophyll a ratio from leaf spectral reflectance. *Photosynthetica*, 31 (2), 221-230.

Penuelas, J., Filella, I., Biel, C., Serrano, L., and Save, R. (1993a). The reflectance at the 950–970 nm region as an indicator of plant water status. *International Journal of Remote Sensing*, 14 (10), 1887-1905.

Penuelas, J., Gamon, J., Griffin, K., and Field, C. (1993b). Assessing community type, plant biomass, pigment composition, and photosynthetic efficiency of aquatic vegetation from spectral reflectance. *Remote Sensing of Environment*, 46 (2), 110-118.

Penuelas, J., Pinol, J., Ogaya, R., and Filella, I. (1997). Estimation of plant water concentration by the reflectance water index WI (R900/R970). *International Journal of Remote Sensing*, 18 (13), 2869-2875.

Peters, A., Hothorn, T., and Lausen, B. (2002). ipred: Improved Predictors. *R news,* 2/2 33-36.

Portigal, F., Holasek, R., Mooradian, G., Owensby, P., Dicksion, M., and Fene, M. (1997). Vegetation classification using red edge first derivative and green peak statistical moment indices with the Advanced Airborne Hyperspectral Imaging System(AAHIS). In

Prasad, A., Iverson, L., and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9 (2), 181-199.

Price, J. (1992). Variability of high-resolution crop reflectance spectra. *International Journal of Remote Sensing*, 13 (14), 2593-2610.

Price, J. (1994). How unique are spectral signatures? *Remote Sensing of Environment*, 49 (3), 181-186.

Prinzie, A., and Van den Poel, D. (2008). Random forests for multiclass classification: Random multinomial logit. *Expert systems with Applications*, 34 (3), 1721-1732.

Proisy, C., Couteron, P., and Fromard, F. (2007). Predicting and mapping mangrove biomass from canopy grain analysis using Fourier-based textural ordination of IKONOS images. *Remote Sensing of Environment*, 109 (3), 379-392.

Pu, R., Yu, Q., Gong, P., and Biging, G. (2005). EO-1 Hyperion, ALI and Landsat 7 ETM+ data comparison for estimating forest crown closure and leaf area index. *International Journal of Remote Sensing*, 26 (3), 457-474.

Qi, J., Moran, M., Cabot, F., and Dedieu, G. (1995). Normalization of sun/view angle effects using spectral albedo-based vegetation indices. *Remote Sensing of Environment*, 52 (3), 207-217.

Questier, F., Put, R., Coomans, D., Walczak, B., and Heyden, Y. (2005). The use of CART and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems*, 76 (1), 45-54.

R Development Core Team (2007). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rahman, A., Gamon, J., Fuentes, D., Roberts, D., and Prentiss, D. (2001). Modeling spatially distributed ecosystem flux of boreal forest using hyperspectral indices from AVIRIS imagery. *Journal of Geophysical Research*, 106 (D24), 33579.

Ramsey, E., and Jensen, J. (1996). Remote sensing of mangrove wetlands: relating canopy spectra to site-specific data. *Photogrammetric engineering and remote sensing*, 62 (8), 939-948.

Ray, S., Das, G., Singh, J., and Panigrahy, S. (2006). Evaluation of hyperspectral indices for LAI estimation and discrimination of potato crop under different irrigation treatments. *International Journal of Remote Sensing*, 27 (23-24), 5373-5388.

Rendonga, L., and Jiyuanb, L. (2004). Estimating wetland vegetation biomass in the Poyang Lake of central China from Landsat ETM data. *IEEE Transactions on Geoscience and Remote Sensing IGARSS apo*, 4 4590-4593.

Richards, J., and Jia, X. (2006). *Remote Sensing Digital Image Analysis: An Introduction*. (Fourth ed.). Springer Verlag.

Ringrose, S., Vanderpost, C., and Matheson, W. (2003). Mapping ecological conditions in the Okavango delta, Botswana using fine and coarse resolution systems including simulated SPOT vegetation imagery. *International Journal of Remote Sensing*, 24 (5), 1029-1052.

Rondeaux, G., Steven, M., and Baret, F. (1996). Optimization of soil-adjusted vegetation indices* 1. *Remote Sensing of Environment*, 55 (2), 95-107.

Rosso, P., Ustin, S., and Hastings, A. (2005). Mapping marshland vegetation of San Francisco Bay, California, using hyperspectral data. *International Journal of Remote Sensing*, 26 (23), 5169-5191.

Rouse, J.W., Haas, R.H., Schell, J.A., and Deering, D.W. (1973). Monitoring vegetation systems in the Great Plains with ERTS. In, *Third ERTS Symposium* (pp. 309 – 317). Washington, DC (NASA), Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., and Harlan, J.C. (1974). Monitoring the vernal advancement and retrogradation of natural vegetation. In (p. 371). NASA/GSFC

Satterwhite, M., and Ponder Henley, J. (1987). Spectral characteristics of selected soils and vegetation in northern Nevada and their discrimination using band ratio techniques. *Remote Sensing of Environment*, 23 (2), 155-175.

Sawaya, K., Olmanson, L., Heinert, N., Brezonik, P., and Bauer, M. (2003). Extending satellite remote sensing to local scales: land and water resource monitoring using high-resolution imagery. *Remote Sensing of Environment*, 88 (1-2), 144-156.

Schlerf, M., Atzberger, C., and Hill, J. (2005). Remote sensing of forest biophysical variables using HyMap imaging spectrometer data. *Remote Sensing of Environment*, 95 (2), 177-194.

Schmidt, K., and Skidmore, A. (2003). Spectral discrimination of vegetation types in a coastal wetland. *Remote Sensing of Environment*, 85 (1), 92-108.

Seher, J., and Tueller, P. (1973). Color aerial photos for marshland. *Photogrammetric Engineering*, 39 (5), 489-499.

Serag, M. (2003). Ecology and biomass production of Cyperus papyrus L. on the Nile bank at Damietta, Egypt. *Journal of Mediterranean Ecology*, 4 15-24.

Serrano, L., Pe uelas, J., and Ustin, S. (2002). Remote sensing of nitrogen and lignin in Mediterranean vegetation from AVIRIS data:: Decomposing biochemical from structural signals. *Remote Sensing of Environment*, 81 (2-3), 355-364.

Sha, Z., Bai, Y., Xie, Y., Yu, M., and Zhang, L. (2008). Using a hybrid fuzzy classifier (HFC) to map typical grassland vegetation in Xilin River Basin, Inner Mongolia, China. *International Journal of Remote Sensing*, 29 (8), 2317-2337.

Shaikh, M., Green, D., and Cross, H. (2001). A remote sensing approach to determine environmental flows for wetlands of the Lower Darling River, New South Wales, Australia. *International Journal of Remote Sensing*, 22 (9), 1737-1751.

Shaw, G., and Manolakis, D. (2002). Signal processing for hyperspectral image exploitation. *IEEE Signal Processing Magazine*, 19 (1), 12-16.

Shen, S.S. (2007). Optimal band selection and utility evaluation for spectral systems. In C.-I. Chang (Ed.), *Hyperspectral Data Exploitation: Theory and Applications* (pp. 227–243). New York, USA. : John Wiley and Sons, Inc.

Shima, L., Anderson, R., and Carter, V. (1976). The use of aerial color infrared photography in mapping the vegetation of a freshwater marsh. *Chesapeake Science*, 17 (2), 74-85.

Silva, T., Costa, M., Melack, J., and Novo, E. (2008). Remote sensing of aquatic vegetation: theory and applications. *Environmental Monitoring and Assessment*, 140 (1), 131-145.

Sims, D., and Gamon, J. (2002). Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sensing of Environment*, 81 (2-3), 337-354.

Skidmore, A., Forbes, G., and Carpenter, D. (1988). Technical note Non-parametric test of overlap in multispectral classification. *International Journal of Remote Sensing*, 9 (4), 777-785.

Smith, K.L., Steven, M.D., and Colls, J.J. (2004). Use of hyperspectral derivative ratios in the red-edge region to identify plant stress responses to gas leaks. *Remote Sensing of Environment*, 92 (2), 207-217.

Soh, L., and Tsatsoulis, C. (1999). Segmentation of satellite imagery of natural scenes using data mining. *IEEE Transactions on Geoscience and Remote Sensing*, 37 (2), 1086-1099.

Steininger, M. (2000). Satellite estimation of tropical secondary forest above-ground biomass: data from Brazil and Bolivia. *International Journal of Remote Sensing*, 21 (6), 1139-1157.

Stimson, H.C., Breshears, D.D., Ustin, S.L., and Kefauver, S.C. (2005). Spectral sensing of foliar water conditions in two co-occurring conifer species: Pinus edulis and Juniperus monosperma. *Remote Sensing of Environment*, 96 (1), 108-118.

Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8 (1), 25.

Suits, G.H. (1983). The nature of electromagnetic radiation. In D.S.S.a.F.T.U. R. N. Colwell (Ed.), *Manual of Remote Sensing : Theory, Instruments and Techniques* (pp. 36-60): Falls Church, Va.: American Society of Photogrammetry.

Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R., and Feuston, B. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Scienci*, 43 (6), 1947-1958.

Swain, P., and Davis, S. (1978). *Remote sensing: the Quantitative Approach*. New York, McGraw-Hill.

Tan, Q., Shao, Y., Yang, S., and Wei, Q. (2003). Wetland vegetation biomass estimation using Landsat-7 ETM+ data. *IEEE Transactions on Geoscience and Remote Sensing, IGARSS apos*, 03 (4), 2629-2631.

Tanriverdi, C. (2006). A Review of remote sensing and vegetation indices in precision farming. *Journal of Science and Engineering*, 9 69-76.

Taylor, R.H. (1995). *St- Lucia Wetland Park*. Cape Town,South Africa: Struik Publishers.

Thenkabail, P., Enclona, E., Ashton, M., and Van Der Meer, B. (2004). Accuracy assessments of hyperspectral waveband performance for vegetation analysis applications. *Remote Sensing of Environment*, 91 (3-4), 354-376.

Thenkabail, P., Smith, R., and De Pauw, E. (2000). Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote Sensing of Environment*, 71 (2), 158-182.

Thenkabail, P., Smith, R., and De Pauw, E. (2002). Evaluation of narrowband and broadband vegetation indices for determining optimal hyperspectral wavebands for agricultural crop characterization. *Photogrammetric engineering and remote sensing*, 68 (6), 607-622.

Thompson, K., Shewry, P., and Woolhouse, H. (1979). Papyrus swamp development in the Upemba Basin, Zaire: studies of population structure in Cyperus papyrus stands. *Botanical Journal of the Linnean Society*, 78 (4), 299-316.

Todd, S., Hoffer, R., and Milchunas, D. (1998). Biomass estimation on grazed and ungrazed rangelands using spectral indices. *International Journal of Remote Sensing*, 19 (3), 427-438.

Toomey, M., and Vierling, L. (2006). Estimating equivalent water thickness in a conifer forest using Landsat TM and ASTER data: a comparison study. *Can. J. Remote Sensing*, 32 (4), 288-299.

Tsai, F., Lin, E., and Yoshino, K. (2007). Spectrally segmented principal component analysis of hyperspectral imagery for mapping invasive plant species. *International Journal of Remote Sensing*, 28 (5-6), 1023-1040.

Tucker, C. (1977). Asymptotic nature of grass canopy spectral reflectance. *Applied Optics*, 16 (5), 1151-1156.

Tucker, C. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8 (2), 127-150.

Ustin, S., Roberts, D., Gamon, J., Asner, G., and Green, R. (2004). Using imaging spectroscopy to study ecosystem processes and properties. *BioScience,* 54 (6), 523-534.

Vaiphasa, C., Ongsomwang, S., Vaiphasa, T., and Skidmore, A. (2005). Tropical mangrove species discrimination using hyperspectral data: A laboratory study. *Estuarine, Coastal and Shelf Science*, 65 (1-2), 371-379.

Vaiphasa, C., Skidmore, A.K., de Boer, W.F., and Vaiphasa, T. (2007). A hyperspectral band selector for plant species discrimination. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62 (3), 225-235.

van Aardt, J., and Norris-Rogers, M. (2008). Spectral-age interactions in managed, even-aged Eucalyptus plantations: application of discriminant analysis and classification and regression trees approaches to hyperspectral data. *International Journal of Remote Sensing*, 29 (6), 1841-1845.

Vogelmann, J., Rock, B., and Moss, D. (1993). Red edge spectral measurements from sugar maple leaves. *International Journal of Remote Sensing*, 14 (8), 1563-1575.

Wang, C., Menenti, M., Stoll, M., Belluco, E., and Marani, M. (2007). Mapping mixed vegetation communities in salt marshes using airborne spectral data. *Remote Sensing of Environment*, 107 (4), 559-570.

Xie, Y., Sha, Z., and Yu, M. (2008). Remote sensing imagery in vegetation mapping: a review. *Journal of Plant Ecology*, 1 (1), 9.

Xu, M., Watanachaturaporn, P., Varshney, P., and Arora, M. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, 97 (3), 322-336.

Xue, L., and Yang, L. (2009). Deriving leaf chlorophyll content of green-leafy vegetables from hyperspectral reflectance. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64 (1), 97-106.

Yang, C., Prasher, S., Enright, P., Madramootoo, C., Burgess, M., Goel, P., and Callum, I. (2003). Application of decision tree technology for image classification using remote sensing data. *Agricultural Systems*, 76 (3), 1101-1117.

Yang, X. (2007). Integrated use of remote sensing and geographic information systems in riparian vegetation delineation and mapping. *International Journal of Remote Sensing*, 28 (2), 353-370.

Yang, Y., Xiao, Y., and Segal, M. (2005). Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics*, 21 (7), 1084.

Yuan, L., and Zhang, L. (2006). Identification of the spectral characteristics of submerged plant Vallisneria spiralis. *Acta Ecologica Sinica*, 26 (4), 1005-1010.

Zarco-Tejada, P., Berjón, A., López-Lozano, R., Miller, J., Martín, P., Cachorro, V., González, M., and De Frutos, A. (2005). Assessing vineyard condition with hyperspectral indices: Leaf and canopy reflectance simulation in a row-structured discontinuous canopy. *Remote Sensing of Environment*, 99 (3), 271-287.

Zarco-Tejada, P.J. (1998). Optical Indices as Bioindicators of Forest Sustainability: Research Evaluation Course. In. Toronto, Canada: York University.

Zhang, H., and Wang, M. (2009). Search for the smallest random forest. *Statistics and its interface*, 2 (3), 381.

Zhang, J., and Foody, G. (1998). A fuzzy classification of sub-urban land cover from remotely sensed imagery. *International Journal of Remote Sensing*, 19 (14), 2721-2738.

Zhao, D., Huang, L., Li, J., and Qi, J. (2007). A comparative analysis of broadband and narrowband derived vegetation indices in predicting LAI and CCD of a cotton canopy. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62 (1), 25-33.

Zheng, D., Rademacher, J., Chen, J., Crow, T., Bresee, M., Le Moine, J., and Ryu, S. (2004). Estimating aboveground biomass using Landsat 7 ETM+ data across a managed landscape in northern Wisconsin, USA. *Remote Sensing of Environment*, 93 (3), 402-411.

Zomer, R., Trabucco, A., and Ustin, S. (2009). Building spectral libraries for wetlands land cover classification and hyperspectral remote sensing. *Journal of environmental management*, 90 (7), 2170-2177.